ABSTRACT

Title of dissertation: USING PHYLOTRANSCRIPTOMICS TO STUDY THE EVOLUTION OF THE GREEN ALGAE

David Anthony Ferranti, Master of Science, 2021

Dissertation directed by: Professor Charles F. Delwiche
Department of Cell Biology and Molecular Genetics

The colonization of land by plants approximately 500 million years ago (Ma) is one of the most important events in the history of complex life. Land plants, hereafter referred to as "embryophytes," comprise the ecological foundation of every major terrestrial biome, making them an essential lineage to the origin and maintenance of biodiversity in those habitats. The embryophytes form a monophyletic clade within one of the two major phyla of the green algae, the charophytes. Estimates from both fossil data and molecular clock analyses suggest that the charophytes diverged from the other main phylum of green algae, the chlorophytes, by as much as 1500 Ma. Here I present a phylogenetic analysis using transcriptomic and genomic data of 62 green algae and embryophyte operational taxonomic units, 31 of which were assembled *de novo* for this project. I focus on identifying the charophyte lineage that is sister to embryophytes, and show that the Zygnematophyceae have the strongest support, although the Charophyceae also have moderate support. I demonstrate that this phylogenetic tree topology is robust across different phylogenetic models and methods. Furthermore, I examine amino acid and codon usage across the tree and find that patterns in these data broadly follow the phylogenetic tree. I conclude by searching my dataset for the presence/absence of several protein domains and gene families known to be important in embryophytes, including the ethylene signaling pathway and various ion transporters. Many of these domains and genes have homologous sequences in the charophyte lineages, indicating that those green algae were particularly well-suited to the colonization of land

USING PHYLOTRANSCRIPTOMICS TO STUDY THE EVOLUTION OF THE GREEN ALGAE

by

David Anthony Ferranti

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2021

Advisory Committee:

Professor Charles Delwiche, Chair
Professor Caren Chang
Professor Stephen Mount

Acknowledgements

I would like to thank Endymion Cooper for generating the bulk of the data used in this thesis project, Charlie Goodman for discussions on bioinformatics, and the rest of the Delwiche lab for their encouragement and support on entering the world of algae research. I would also like to thank our collaborators for providing data: the Mary Bisson lab at the University of Buffalo, the Jaakko Hyvönen lab at the University of Helsinki, and the Fay-Wei Li lab at Cornell University. Special thanks to Thomas Doak and the National Center for Genome Analysis Support (NCGAS) at the University of Indiana for providing computational resources and Heven Sze at the University of Maryland College Park for providing expertise on plant biology.

Stephen Mount and Caren Chang, my committee members, both contributed immensely to the analyses presented in this document and were instrumental in producing a final thesis project under unusual circumstances.

Finally, I would like to thank Charles Delwiche for his mentorship, support, and training in how to approach evolutionary questions with rigor, insight, and humility.

Table of Contents

Table of Abbreviations

BUSCO	Benchmarking Universal Single-Copy Orthologs
C; Chlorodendro	Chlorodendrophyceae
Cha; Charo	Charophyceae s. s.
Chloro	Chlorophyceae
Col; Coleo	Coleochaetophyceae
Early; Early Vas	Early Vascular
HMM	Hidden Markov model
JTT	Amino acid substitution matrix (Jones *et al.* 1992)
K; Klebsormid	Klebsormidiophyceae
LG	Amino acid substitution matrix (Le and Gascuel 2008)
M; Meso	Mesostigmatophyceae
P; Pras	Prasinophytina
Pre; Pre-Vas	Pre-Vascular
T; Trebouxio	Trebouxiophyceae
Ulvo	Ulvophyceae
WAG	Amino acid substitution matrix (Whelan and Goldman 2001)
Zyg; Zygnemato	Zygnematophyceae

Introduction

The colonization of land by plants approximately 500 Ma is one of the most important events in the history of complex life (Heckman *et al.* 2001). Land plants, hereafter referred to as "embryophytes," comprise the ecological foundation of every major terrestrial biome, making them an essential lineage to the origin and maintenance of biodiversity in those habitats. Embryophytes are also foundational crop species and thus critical to the development of complex human societies. Additionally, many embryophytes also produce important products for global commerce and have medicinal applications. The embryophytes form a monophyletic clade within one of the two major phyla of the green algae, the "charophytes s. l." (sensu lato) (Delwiche and Cooper 2015; Delwiche and Timme 2011; Mattox and Stewart 1984). For the purposes of this manuscript, charophytes s. l. refers to the entire charophyte lineage and all its descendants, including embryophytes. Charophytes s. s. (sensu stricto) refers only to the lineage including *Chara, Nitella, Tolypella*, and the other organisms known as stoneworts. It should be noted that the fossil diversity of Charophytes s.s. greatly exceeds their extant diversity, but that only a minimal fossil record exists for other non-embryophyte charophytes (Tappan 1980).

Estimates from both fossil data and molecular clock analyses suggest that the charophytes s. l. diverged from the other main phylum of green algae, the "chlorophytes," by as much as 1500 Ma (Del Cortona *et al.* 2020; Leliaert *et al.* 2011; Hedges *et al.* 2004). Together, the charophytes s. l. and the chlorophytes comprise the "Chloroplastida" clade (Adl *et al.* 2019). Charophyte algae consist entirely of freshwater organisms, although a few taxa have adapted to persist in subaerial and saline environments, whereas chlorophytes, although likely freshwater in origin, are found in marine, freshwater, and terrestrial environments, having independently colonized land several times (Fučíková *et al.* 2014; Blank 2013; Leliaert *et al.* 2012). Most studies focused on

characterizing chlorophyte diversity and evolutionary relationships have dealt with the "core chlorophyte" clades (Chlorophyceae, Trebouxiophyceae, and Ulvophyceae), although there have been efforts in recent years to examine the more basal chlorophytes, including the deep-water marine Palmophyllophyceae and the various prasinophyte clades (Leliaert *et al.* 2016). The Trebouxiophyceae are typically thought to be the outgroup to the remaining two core chlorophyte clades, although it has been proposed that neither Trebouxiophyceae nor Ulvophyceae are truly monophyletic, and a new classification system is needed to properly describe the core chlorophyte clades (Del Cortona *et al.* 2020; Fučíková *et al.* 2014). Much of the work on the evolution of the charophytes s. l. relates to the placement of different charophyte lineages in relationship to both each other and the ancestor to all modern land plants. In particular, the identification of the charophyte s. l. lineage that is the sister to embryophytes has been an object of study for decades (Lewis and McCourt 2004; Turmel *et al.* 2003; Karol *et al.* 2001; Bhattacharya and Medlin 1998; Graham 1996; Mishler and Churchill 1985). Understanding the evolutionary relationship between the charophytes s. l. and embryophytes as well as the features that allowed for the algal transition onto land is important to advancing our knowledge of how embryophytes were able to persist in and later dominate all terrestrial ecosystems (Bowles *et al.* 2020; Cheng *et al.* 2019; Harholt *et al.* 2016).

Three separate charophyte algal clades are prominent candidates for the sister taxon to embryophytes: the Charophyceae s. s., the Coleochaetophyceae, and the Zygnematophyceae. Taxa in all three clades display complex morphology, sexual reproduction, and multicellularity, although there are some members of the Zynematophyceae, known as desmids, that are single-celled, implying that multicellularity was present in the ancestor of those three clades and thereafter lost along the branch leading to the desmid lineage (Wicket *et al.* 2014; but see Cheng

et al. 2019). Prior to the advent of constructing phylogenies from molecular sequence data, these features made the Charophyceae s. s., the Coleochaetophyceae, and the Zygnematophyceae the primary candidates under consideration for the immediate outgroup to embryophytes (Wicket *et al.* 2014). Since then, separate phylogenetic analyses using a combination of plastid, mitochondrial, and nuclear genes have recovered support for each clade being the true sister taxon (Finet *et al.* 2010; but see Laurin-Lemay *et al.* 2012; Turmel *et al.* 2006; Turmel *et al.* 2003; McCourt *et al.* 2004; Karol *et al.* 2001; Mishler *et al.* 1994). Although the most current phylogenomic studies indicate that the Zygnematophyceae are the sister lineage, additional work with denser taxon sampling from the different green algal lineages, chlorophytes and charophytes s. l. alike, is needed to resolve the topology of the Chloroplastida phylogenetic tree where the charophytes s. l. give way to early embryophytes such as the hornworts, liverworts, and mosses, collectively known as bryophytes (Wickett *et al.* 2014; Timme *et al.* 2012). Previous analyses have significantly advanced our understanding of the transition of the charophyte algae onto land, but often relied on either only a few algal taxa or a handful of genes, both of which can hinder phylogenetic analyses that seek to determine deep evolutionary relationships.

Recent efforts by researchers around the world have created the conditions to re-examine the green algae phylogeny with large-scale datasets. The advent of high-throughput sequencing data has allowed for the development of many high-quality omics resources for many non-model organisms, green algae and early-diverging embryophytes among them (Jiao *et al.* 2020; Cheng *et al.* 2019; Nishiyama *et al.* 2018; Delwiche *et al*. 2017; Hori *et al.* 2014). Both transcriptomic and genomic data are powerful in constructing phylogenies using multiple sequence alignments from dozens or even hundreds of genes. These large alignments are key to reconstructing deep

evolutionary relationships, as they give sufficient power to the maximum-likelihood and Bayesian methods necessary for complex phylogenetic construction and the testing of evolutionary hypotheses (Cheon *et al.* 2020; McKain *et al.* 2018; Wickett *et al.* 2014). Due to the high amount of species diversity and long divergence times found across Chloroplastida, it has been difficult to answer questions that span the evolution of the entire clade. Even merely identifying orthologous sequences from a sufficient number of taxa from across the tree for alignment and additional analysis can be difficult since various Chloroplastida clades, especially embryophytes, have undergone rapid genome evolution (Proost *et al.* 2011). Furthermore, phylogenetic analysis may become inconsistent and generate artifacts when highly diverged taxa with many accumulated character changes are incorrectly grouped together. Long branch attraction is therefore more likely to occur when analyzing multiple clades that are distantly related and presents a challenge when attempting to answer evolutionary questions that span hundreds of millions of years (Brinkmann *et al.* 2005). Recent advancements in phylogenetic methods, including models that attempt to deal with heterotachous sequence evolution, and efforts at increasing the breadth and width of taxon sampling, however, have proven useful in increasing the power of phylogenetic analysis to test existing hypotheses about the evolution of clades across Chloroplastida, including the Ulvophyceae green seaweeds, early-diverging charophytes, bryophytes, and gymnosperms (Del Cortona *et al.* 2020; Cox *et al.* 2014; Zhong *et al.* 2014; Rodríguez-Ezpeleta *et al.* 2007).

The sequencing and assembly of the *Chara braunii* genome, as well as the genomes of several species of Zygnematophyceae, have provided valuable resources for investigation of the sister lineage to embryophytes. Furthermore, omics analyses and manipulative laboratory experiments performed on various Zygnematophyceae taxa have shown groups of orthologous

genes associated with both biotic and abiotic stress in embryophytes as well as similar cellular responses to desiccation and intense light stress (Jiao *et al.* 2020; De Vries *et al.* 2018; Holzinger and Pichrtová 2017). Since bioinformatic approaches can play a role in adding important evolutionary context as well as suggesting candidate genes and pathways for further investigation, they make for an excellent complement to molecular biology work. Examining the presence of protein domains across Chloroplastida classified by both phylogenetic clade and freshwater, marine, and terrestrial habitats presents another opportunity to interrogate the genomic features of the colonization of land. It has been suggested that the charophyte algae were uniquely suited to survive in and eventually exploit terrestrial habitats due to their genomic and molecular toolkit, although to date no single gene or metabolic capability has been identified that is key to the transition. Rather, it seems that a complex combination of attributes was required (De Vries *et al.* 2018; Delaux *et al.* 2015). Previous work has shown that homologous sequences to genes involved in the ethylene signaling pathway in embryophytes are present in several charophyte algal lineages but absent in chlorophytes. Additionally, *Spirogyra*, a filamentous Zygnematophyceae charophyte taxa, was shown to produce ethylene and use the hormone to regulate the elongation of cells, providing an important clue about the potential utility of plant pathways in the likely ancestral algal lineage (Ju *et al.* 2015). Mapping the presence of different components of a modern-day organism's genomic toolkit throughout a phylogenetic tree is a useful exploratory method for generating hypotheses about the emergence of important biological pathways. Moreover, it provides a method to detect possible cases of convergent evolution, wherein distantly related clades independently acquire similar proteins or protein domains in response to similar environmental pressures such as desiccation stress or UV

light stress. Both of these stressors are likely to have played a role in shaping the algal colonization of land (Jiao *et al.* 2020).

Here I present a maximum-likelihood phylogenetic analysis using an alignment built from more than a thousand orthologous sequences of 62 operational taxonomic units that span the Chloroplastida tree of life from early-diverging marine prasinophytes to angiosperms. I focus on identifying the charophyte algal lineage that is sister to embryophytes by testing a series of alternative phylogenetic tree topologies that shuffle the positions of the Zygnematophyceae, Coleochaetophyceae, and Charophyceae s. s. clades in relation to both embryophytes and the older green algae clades. I further examine how both amino acid and codon usage vary across the phylogenetic tree as well as quantifying the number of protein domains that are shared between different major lineages of Chloroplastida and specific lineages of interest within the green algae. I conclude by tracing the presence of several protein domains found in the ethylene signaling pathway as well as several ion transporter gene families through the different algal lineages.

Methods

De novo *assemblies and dataset construction*

I used 32 paired-end RNA-Seq datasets originally prepared by Endymion Cooper, a former postdoctoral researcher in the Delwiche lab. He extracted poly-A selected mRNA from 31 species of green algae and sequenced it high-throughput shotgun sequencing on an Illumina machine (Table 1). Two libraries were prepared for *Nitella mirabilis,* one each from the upper and lower portions of the organism. A single library was prepared for all other taxa. Reads were de-multiplexed using the Illumina Casava pipeline. To remove poor-quality reads, poly-A tails, and singletons, reads were then lightly filtered with PRINSEQ-lite version 0.20.4 running on the following settings: -trim_qual_left 20, -trim_qual_right 20, -trim_qual_window 20, -trim_tail_left 101, -trim_tail_right 101, -trim_ns_left 1, -trim_ns_right 1, -min_len 25, -min_qual_mean 20, and -out_format 3 (Schmieder and Edwards 2011).

I began my analyses with this set of filtered reads. The filtered reads were then assembled using rnaSPAdes 3.14.0 with default options (Bushmanova *et al.* 2019). In the case of *Nitella mirabilis*, both pairs of reads, from the libraries prepared from the upper and lower parts of the plant, were used to assemble the transcriptome. Assembled transcriptomes were filtered by running blastx searches of each assembled transcript against the NCBI nr.fasta database (downloaded June 2020) using DIAMOND BLAST 2.0.4 (Buchfink *et al.* 2014). Due to suspected fungal contamination, transcripts were only retained if the best BLAST hit they produced had a bitscore above 200 against a sequence labeled as *Viridiplantae.* All other transcripts were discarded. Assemblies were converted into amino acid sequences with TransDecoder v5.5.0 and assessed for completeness by running BUSCO v4.0.6 in protein mode

against the Chlorophyte, Viridiplantae, and Eukaryote databases (Haas *et al.* 2013; Seppey *et al.* 2019).

Data for other taxa were either downloaded from publicly available sources or provided by collaborators as transcriptomes or coding sequences from the genome (Table 2). For both types of data, the nucleotide sequences were converted into amino acid sequences with TransDecoder v5.5.0 and assessed for completeness using BUSCO as with the *de novo* transcriptome assemblies described above.

Ortholog identification and multiple sequence alignment

Due to the large divergence time of the taxa in our dataset, traditional reciprocal BLAST search methods failed to identify a sufficient number of orthologous genes for phylogeny construction. Consequently, I used a hidden Markov model approach to identify orthologs in each taxa for multiple sequence alignment. First, coding sequences were obtained from nine (three charophyte and six chlorophyte) different green algae genome assemblies (Table 3). The coding sequences were checked for completeness with BUSCO and then converted into amino acid sequences with TransDecoder. Orthofinder was used to identify orthogroups in those nine algae (Emms and Kelly 2015; Emms and Kelly 2018). Orthogroups that contained at least one sequence from each of the nine algae species (hereafter referred to as universal orthogroups) were aligned using MAFFT v7.471 and used to construct hidden Markov models using HMMER3.1 (Katoh *et al.* 2002; Eddy 2009). Orthologs were then identified in each of the species in the larger dataset by searching the species' assembly with the hidden Markov model generated from each universal orthogroup and extracting the sequence that produced the best hit according to bitscore. The *Chara braunii* and *Penium margaritaceum* genomes were also used in the later phylogenetic

analyses in order to increase the sampling of the Charophyceae s. s. and Zygnematophyceae respectively.

Phylogenetic analyses

Individual orthologs were aligned using MAFFT v7.471 and then concatenated into a single superalignment. To reduce the percentage of gaps in the alignment, trimAl v1.3 was used to remove sites in the superalignment that contained a gap for 25% or more taxa (Capella-Gutiérrez *et al.* 2009). IQ-TREE v1.6.12 was used to generate a maximum likelihood phylogeny using a model of sequence evolution consisting of the LG substitution matrix, empirical base frequencies, and ten rate parameters (LG+F+R10) and ten thousand ultrafast bootstrap replicates (Nguyen *et al.* 2015; Hoang *et al.* 2017). The model was selected by using IQ-TREE's inbuilt ModelFinder option to calculate the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for each of 78 total protein models using the JTT, LG, and WAG substitution matrices and up to ten rate parameters (Kalyaanamoorthy *et al.* 2017; Le and Gascuel 2008; Whelan and Goldman 2001; Yang 1995; Jones *et al.* 1992). I also performed an analysis using the GHOST model with four categories (LG+F*H4) to attempt to account for heterotachous sequence evolution (Crotty *et al.* 2020). Lastly, I used ModelFinder to calculate the optimal amino acid matrix (JTT, LG, or WAG) and number of rate parameters for each of the individual trimAl filtered alignments for the 1323 orthologous genes and ran an edge-unlinked partitioned analysis with each individual gene using its optimal model. The same individual gene trees calculated from the analysis of each individual orthologous gene were fed into ASTRAL III to produce a reconciliation tree (Zhang *et al.* 2018).

I further investigated the topology of the phylogenetic tree related to the algal colonization of land by using likelihood mapping analysis and the groupings shown in Table 4 (Strimmer and

Von Haeseler 1997). I performed a total of seven likelihood mapping analyses. Three of the analyses ignored the "Older Taxa" grouping and calculated the support under the LG+F+R10 model, the edge-unlinked partition model, and the GHOST heterotachy model. The other four likelihood mapping analyses were performed with the LG+F+R10 model and moved the Older Taxa grouping to be included in each of the Charophyceae s. s., Coleochaetophyceae, Zygnematophyceae, and Embryophyte groupings. In each analysis, I recorded the percentage of likelihood quartets that showed each of the three charophyte algae clades of interest as the sister taxa of Embryophytes in the resulting four-taxon tree. In addition to the various likelihood mapping analyses, I also calculated the likelihood (under the LG+F+R10 model) of each of the 15 possible 5-taxon tree topologies consisting of the separate Embryophyte, Charophyceae s. s., Coleochaetophyceae, Zygnematophyceae, and Older Taxa groupings.

One final phylogenetic analysis was performed using the LG+F+R10 model and a different multiple sequence alignment originally created by Endymion Cooper. This alignment consisted of approximately 1700 orthologous genes across 35 species.

Amino acid and codon composition analyses

I calculated the frequency of each amino acid across the orthologous genes for each species and performed a principal components analysis on the data. I then calculated amino acid frequencies for each species' entire protein assembly as well as codon frequencies for the corresponding nucleotide coding sequences and ran a principal components analysis for both sets of those frequency data as well. Analyses were performed and visualized using the FactoMineR and factoextra packages in R (Lê *et al.* 2008; Kassambara 2016).

Conserved domain database searches

I used the Conserved Domain Database (CDD) from NCBI to search the combined Chloroplastida dataset, including the species used in universal orthogroup identification and the phylogenetic analyses, for patterns of protein domain presence and absence. Nineteen thousand two hundred and seven protein domain multiple sequence alignments were downloaded in winter 2021 and used to build hidden Markov models. HMMER searches of each model were performed against the combined protein database of the algal genomes used to identify orthologs and the combined genomic and transcriptomic data included in the phylogenetic analyses. Domains were classified as being present or absent in each of the major clade and habitat classifications given in tables 1, 2, and 3. Domains were considered present in a clade if at least one species in the clade produced a hit with an E-value less than $1*10^{-51}$ against the hidden Markov model built from the multiple sequence alignment of the domain. All other hits were discarded.

Individual domain and protein family searches

I also searched for a variety of specific protein domains found in genes involved in the ethylene signaling pathway in land plants, including CTR1, EIN2, EIN3, and ETR1. Searches were performed in DIAMOND BLAST with the amino acid sequence of each domain from the corresponding *Arabidopsis thaliana* gene as the query against the combined protein database of the algal genomes used to identify orthologs and the combined genomic and transcriptomic included in the phylogenetic analyses. Hits to *Arabidopsis thaliana* were excluded to avoid distortion of the results. Specific domains in the search include the N-terminal domain of CTR1, the signaling domain of EIN2, and the ethylene-binding domain of ETR1. Since there is no distinguishing protein domain in the sequence coded for by EIN3, the entire sequence was searched. Searches were also made for the subdomains of the signaling domain in EIN2

previously identified by John Clay's dissertation in *Spirogyra, Nitella, and Chlamydomonas.* Gene family searches were performed for the sodium/hydrogen antiporter (NHX), cation/hydrogen exchanger (CHX), potassium efflux antiporter (KEA), and amino acid permease (AAP) gene families. As before, searches were performed with DIAMOND BLAST using *Arabidopsis thaliana* sequences against the combined databases described above, and hits to *Arabidopsis thaliana* were removed before visualization. 6 NHX genes, 4 CHX genes, 6 KEA genes, and 8 AAP genes were searched for a total of 24 genes from all families. Both the *Arabidopsis* AAP sequences and sequences in the database that produced a BLAST hit with an information density (bitscore divided by query length) greater than 0.75 from the AAP search were extracted and used to build a phylogenetic tree. Additional sequences added to this tree from GenBank include a *Trebouxia* sequence, a diatom (*Fragilariopsis cylindrus*) sequence, a fungal (*Mortierella*) sequence, and two bacterial sequences (*Rhizobium*).

Results

Dataset quality and completeness

Mean BUSCO completeness scores for each species included in the phylogenetic analysis compared against the Chlorophyte, Eukaryote, and Viridiplantae databases are shown in Figure 1. Taxa are organized by the "Minor Clade" classification given in Tables 1 and 2. BUSCO completeness scores for each algal species used to construct the hidden Markov models for ortholog identification are shown in Figure 2. Taxa are organized by the "Major Clade" classification given in Table 3. Presence and absence matrices for each taxa and each BUSCO database can be viewed in Appendix A, and the BUSCO ID and associated function of each BUSCO gene missing from every member of each of the three major clades can be viewed in Appendix B. Since my phylogenetic analyses are reliant on a bioinformatic dataset drawn from many different sources, there will be differences between taxa originating in different extraction protocols, filtration cutoffs, and methods of assembly. Even within closely related species from my own *de novo* assemblies there are large differences. For example, *Chaetosphaeridium globosum* has lower BUSCO scores and a higher percentage of gaps in the multiple sequence alignment than *Coleochaete orbicularis*. The *de novo* transcriptomes were also built using RNA extracted from vegetative tissue, and thus may have failed to include genes involved in specific biological functions such as stress response and reproduction.

OrthoFinder identified a total of 2619 universal orthogroups from the nine algal genomes. The presence/absence of each universal orthogroup in each taxa is shown in Figure 3. After hidden Markov model construction, 1327 of these orthogroups produced a hit for all taxa used in the phylogenetic analysis. After concatenation, TrimAl filtering reduced this number to a final total of 1323 orthologous genes and 566428 sites. The percentage of gap characters for each species

included in the phylogenetic analysis after the trimAl filtration step is shown in Figure 4. Both the *Nothoceros* taxa (hornworts) and multiple gymnosperm taxa show a relatively high percentage of gaps compared to the other taxa in the alignment. No green algae species is composed of greater than 20 % gaps, even those with relatively low BUSCO scores, such as *Nephroselmis pyriformis, Ankistrodemus falcatus,* and *Cephaleuros parasiticus*. In contrast, embryophyte taxa in the alignment with low BUSCO scores, such as the two hornwort *Nothoceros* taxa and most of the gymnosperm taxa *sans Welwitschia*, have the highest percentage of gaps, even above 40%. Nevertheless, they were retained within the analyses in order to represent a sufficiently broad sample of embryophyte taxa. This high percentage of gaps amongst the more incomplete embryophytes may be a consequence of building the HMMs used in orthologous sequence identification from orthogroups that were constructed using only genomes of the green algae. These genomes also suffer from a disadvantage of not capturing the total spectrum of green algae diversity, especially among the chlorophytes. Of all the minor clades within the chlorophytes, only taxa from the Chlorophyceae were included, as no Trebouxiophyceae or Ulvophyceae genomes were available to incorporate into the analysis. Nevertheless, the method was effective at identifying orthologous sequences in all species. Of the 2619 universal orthogroups searched against the species in the phylogenetic analysis, just over half (1323) were found in all species. Here it should also be noted that the *Penium margaritaceum* and *Chara braunii* genomes were used in both HMM construction and in the later phylogenetic analysis, which explains why every universal orthogroup was present in those two datasets. These taxa were included in the phylogenetic analysis in order to increase the sample size of charophytes, particularly the two charophyte clades (Charophyceae s. s. and Zygnematophyceae) that are of interest.

Phylogenetic analyses

The phylogenetic tree from the LG+F+R10 phylogenetic analysis is displayed in Figure 5. The phylogenetic tree from the LG+F*H4 heterotachy analysis is displayed in Figure 6. The distribution of models selected for each individual gene tree is shown in Figure 7, and the phylogenetic tree from the partitioned analysis is displayed in Figure 8. The phylogenetic tree from the ASTRAL gene tree reconciliation analysis is shown in Figure 9. Bootstrap support is shown at each node in each of the trees constructed in IQ-TREE; nodes in the ASTRAL tree show local posterior probability support.

The likelihood analyses that ignored the older taxa resulted in a total of 1680 likelihood quartets. The percentages of likelihood quartets that support the Charophyceae s. s., Coleochaetophyceae, and Zygnematophyceae as the sister lineage to embryophytes are shown in Figure 10. The ranking of the 15 possible five-taxon trees is shown in Figure 11. The phylogenetic tree from Endymion Cooper's alignment is shown in Appendix C. All of the phylogenetic analyses are generally in agreement on the overall topology of the tree, with only minor disagreements on specific branches. Of the three charophyte clades proposed thus far as sister to embryophytes, I recover strong and consistent support for Zygnematophyceae as the true sister lineage to land plants from the various phylogenetic analyses, including the GHOST mixture model analysis, the edge-unlinked partition analysis, the gene tree reconciliation analysis as performed by ASTRAL, and the analysis with Endymion Cooper's alignment.

Amino acid and codon composition analyses

The first, second, and third dimensions of the principal components analysis of amino acid frequencies calculated from orthologous sequences for species used in phylogenetic tree

construction are shown in Figure 12. The first dimension explains 45.3% of the variation in the data, the second dimension explains 15.2% of the variation in the data, the third dimension explains 10.6% of the variation in the data. The first, second, and third dimensions of the principal components analysis of amino acid frequencies calculated from the entire protein assemblies for all species are shown in Figure 13. The first dimension explains 49.3% of the variation in the data, the second dimension explains 19.2% of the variation in the data, and the third dimension explains 15.4% of the variation in the data. The first, second, third, and fourth dimensions of the principal components analysis of codon frequencies calculated from the nucleotide sequences corresponding to the protein assemblies are shown in Figure 14. The first dimension explains 56.3% of the variation in the data, the second dimension explains 13.7% of the variation in the data, and the third dimension explains 10.8% of the variation in the data.

The first dimension of all three principal components analyses explains around half of the variation in the data, indicating that there is structure in the amino acid and codon frequencies among the Chloroplastida. In all three plots, the first principal component generally represents a gradual movement along the phylogeny from chlorophyte to charophyte to embryophyte. *Trentepohlia* and *Cephaleuros* cluster together far away from other chlorophyte in these analyses, likely due to their use of an alternate genetic nuclear code, in which the UAA and UAG codons code for glutamine rather than being treated as stop codons as they are in all other taxa. The major difference between the amino acid data and codon data is found in dimension 2. In the amino acid data, the Trentepohliaes are positioned opposite of the two *Ostreococcus* taxa, but in the codon data they are much closer together. In the amino acid analysis that includes the entire protein assembly, the three grasses (*Oryza, Sorghum,* and *Zea*) in the dataset are closer to the green algae than other embryophytes along the first dimension of the data, but in the amino

acid analysis comprised of only orthologous genes the grasses are grouped with the other embryophytes.

Conserved domain database searches

The conserved domain database hidden Markov model searches resulted in a total of 5883 unique domains that yielded a hit with an E-value less than 10^{-51} across all species in the dataset. Domains unique to chlorophytes, charophytes, and embryophytes, as well as the number of domains shared between each clade, are shown in Figure 15. The conserved domain database entries that are only present in each pairing of the three clades (embryophytes and charophytes, charophytes and chlorophytes, and charophytes and chlorophytes) are listed in Appendix G. Embryophytes have the most unique domains of the three clades. Of the three possible pairings, embryophytes and charophytes share the highest number of domains, which is consistent with both their placement in the phylogeny and with the hypothesis that charophytes possess some of the embryophyte genomic toolkit.

Individual domain and protein family searches

The ethylene signaling pathway protein domain searches are shown in Figures 16-19. Among the algae, the CTR1 N-domain has the strongest signal in the Klebsormidiophyceae and Zygnematoyphceae, the EIN2 signaling domain has the strongest signal in the Coleochaetophyceae and Zygnematophyceae, the EIN3 gene is equally strong in all charophyte lineages past the Mesostigmatophyceae, and the ETR1 ethylene binding domain has the strongest signal in the Zygnematophyceae. The EIN2 subdomain searches for sequences from *Arabidopsis, Spirogyra, Nitella, and Chlamydomonas* are shown in Appendix F. These subdomain searches are more difficult to interpret but generally show homologous sequences to the second and third

subdomains of the EIN2 signaling domain cluster in the Charophyceae s. s., Coleochaetophyceae, Zygnematophyceae, and Chlorophyceae clades, corroborating John Clay's earlier work. The RuBisCO search, which indicates the presence of a gene found in the chloroplast of all the organisms in the tree, is shown in Appendix E and displays scattered hits across the dataset. This result is likely due to different research groups choosing to exclude organellar genomes from their data. My *de novo* assemblies, which were not filtered to remove organellar genes, generally show that RuBisCO is present. The NHX, CHX, KEA, and AAP gene family searches are shown in Figures 20-23. The NHX family has a stronger signal in the charophytes than the chlorophytes and the CHX family is entirely absent from the chlorophytes. The KEA family appears to be universally present through the tree. The AAP family is absent from almost all of the algae, apart from two taxa on a branch in the Trebouxiophytes. The phylogenetic tree of AAP sequences is shown in Appendix D.

Discussion

Green algae phylogenetics and the sister lineage to land plants

My phylogenetic analyses were successful at generating a green algae and broader Chloroplastida phylogeny that is consistent with previously published work by utilizing a dataset comprised of a large number of orthologous genes identified from both genomic and transcriptomic datasets, emphasizing the power of bioinformatic approaches in recovering deep evolutionary relationships (Cheon *et al.* 2018). Although recent multi-gene analyses show Ulvophyceae as being non-monophyletic, including placing *Codium fragile* as more closely related to the Chlorophyceae, I recover strong support for three monophyletic "core chlorophyte" clades in the IQ-TREE analyses (Trebouxiophyceae, Ulvophyceae, and Chlorophyceae) (Del Cortona *et al.* 2020; Fučíková *et al.* 2014). The ASTRAL analysis, however, places *Codium* as the outgroup to all the Chlorophyceae. There are several possible explanations for this discrepancy. First, my analyses may be hindered by a comparative lack of taxon sampling among the Ulvophyceae. Additionally, the algal genomes used in HMM construction for ortholog identification all came from freshwater lineages (Chlorophyceae and various charophyte clades), which may have lessened their ability to detect genes suited for phylogenetic inference within the Ulvophyceae, many of which are green seaweeds and subaerial terrestrial organisms. Since I do not have as many taxa in my analysis as the Del Cortona paper, I am inclined to defer to their tree topology. I recommend additional focus on the Ulvophyceae and Chlorophyceae to parse those evolutionary relationships. My analysis does corroborate the Del Cortona paper's placement of the Trebouxiophyceae as the outgroup to the other two "core chlorophyte" clades. Our analyses also agree on *Mesostigma* and *Chlorokybus* forming a monophyletic clade as opposed to individual early-diverging charophyte lineages. These results corroborate those of

the most recent phylogenomic studies regarding the topology of the Chloroplastida tree (Puttick *et al.* 2014; Wickett *et al.* 2014; Timme *et al.* 2012). Regarding the other two charophyte clades proposed as the sister taxa to embryophytes, the likelihood mapping analyses identify the Charophyceae s. s. as having moderate support for being placed sister to embryophytes. The Coleochaetophyceae have the weakest support but show consistent affinity for being grouped near the Zygnematophyceae in the analysis of the 15 5-taxon trees. In contrast, the Charophyceae s. s. show an affinity for being grouped with both embryophytes and earlier charophytes and chlorophytes, which is curious given that those lineages are highly diverged from one another. This result is consistent with earlier phylogenetic work based on a smaller number of genes that showed the Charophyceae s. s. as the sister lineage to land plants (McCourt *et al.* 2004; Karol *et al.* 2001). It also demonstrates the need for additional work in the Charophyceae s. s., Coleochaetophyceae, and Zygnematophyceae in order to determine their exact position in relation to embryophytes in the topology of the phylogenetic tree.

Insights into algal colonization of terrestrial environments

There is striking diversity of both amino acid and codon usage to be found in the Zygnematophyceae charophyte algae, with *Mougeotia* and the two *Spirogyra*, which are conjugating filamentous algae, clustering closer to embryophytes rather than with the remainder of the green algae, while the single-celled "desmids" (also conjugating green algae) had amino acid and codon usage more characteristic of the bulk of the green algae, both chlorophytes and charophytes. This feature of the Zygnematophyceae does not reflect the phylogeny, with some of the single-celled species in the study being phylogenetically closer to the filamentous species than the other single-celled species. Additionally, *Nitella mirabilis* and the three *Chara* are closer to embryophytes on the first principal component than single-celled Zygnemtaophyceae

(desmids) and the older charophyte clades Mesostigmatophyceae and Klebsormidiophyceae. I advise caution in overinterpreting the any particular principal component from the three analyses due to the varying quality and completeness in the underlying data. Nevertheless, there are some patterns that undoubtedly speak to the underlying biology of the organisms in the analysis. For example, I hypothesize that Charophyceae s. s. and early embryophyte taxa have experienced convergent evolution towards a lifestyle in subaerial terrestrial environments, resulting in a more similar amino acid and codon usage profile. This may also explain the moderate support for the placement of the Charophyceae s. s. as the sister lineage to land plants in the likelihood mapping analyses. I further hypothesize that, in addition to their unusual genetic code, the independent colonization of land by the Trentepohliaes has caused a shift in their use of different amino acids and codons compared to algae living primarily in marine and freshwater habitats. Furthermore, patterns of amino acid and codon usage provide important context for interpreting phylogenetic trees, as biases in codon composition are known to hinder accurate tree construction, as well as giving insight into the nature of large transcriptomic and genomic datasets (Cox *et al. 2014).*

There are two likely reasons for embryophytes having the highest number of unique protein domains in the conserved domain database search. The first reason is ascertainment bias. Since there are many more embryophytes developed as model organisms, including *Arabidopsis thaliana* and several crop species, the multiple sequence alignments of domains in the conserved domain database are more likely to have been constructed from embryophyte sequences. The second reason is the high amount of biological diversity within embryophytes, which has facilitated their successful exploitation of nearly every terrestrial biome. The number of domains shared between clades is also informative. As expected, charophyte taxa share a much higher number of protein domains with embryophyte taxa than charophytes share with chlorophytes or

embryophytes share with chlorophytes. This result emphasizes that the charophyte algal lineages possessed at least some portion of the genomic toolkit that would allow them to persist on land. As charophytes consist solely of freshwater taxa, this result also demonstrates that land plants arose from a strictly freshwater lineage that was uniquely suited to the ensuing colonization of terrestrial habitats (de Vries *et al.* 2018). Recent laboratory manipulations of several Zygnematophyceae taxa have provided empirical evidence that responses to stresses such as high amounts of UV light, which would have been common during the terrestrial transition, are present in charophyte taxa (Jiao *et al.* 2020). Significantly, the phylogenetic analyses point towards Zynematophyceae as the most likely candidate for the sister lineage to land plants, indicating that those particular algae were well-suited to thrive in the available niche space to the ancestral land plant. I recommend that further work, using both experimental and bioinformatic methods, be applied to the charophyte lineages to investigate to what extent they share a genomic toolkit with embryophytes, particularly bryophytes. Generation of additional genomic resources for the charophyte algae, particularly the Charophyceae s. s. and the Zygnematophyceae, would help in disentangling this question. It should also be noted that all three major Chloroplastida clades share most of the protein domains identified in the hidden Markov model search, likely a legacy of their common photosynthetic lifestyle.

My ethylene pathway protein domain searches typically agree with previous work on the evolution of the ethylene signaling pathway (Ju *et al.* 2015). All of the domains (the CTR1 N-terminal domain, the EIN2 signaling domain, and the ETR1 ethylene binding domain), as well as the full sequence of the EIN3 gene, are absent in most of the chlorophyte lineages and show homology to sequences in various charophyte algae beginning with the Klebsormidiophyceae. These hits are particularly strong in the Zygnematophyceae. In the context of previous

experimental manipulations of *Spirogyra,* this seems to indicate that while different components of the ethylene signaling pathway were present in many different groups of the charophyte algae, they may not have had the same function as they do in embryophytes until the Zygnematophyceae. This further reinforces support for the hypothesis that the Zygnematophyceae are the sister lineage to land plants. The comparatively weak hits for the ETR1 ethylene-binding domain in the Charophyceae s. s. and Coleochaetophyceae may be a result of poor data quality, or they may indicate loss of that particular gene and/or domain in those lineages. Broader sampling in the Charophyceae and genomic resources for the Coleochaetophyceae will help resolve this question.

The four ion transporter gene families show distinct patterns across the Chloroplastida phylogeny. First, the NHX gene family is broadly present in all clades, but charophyte sequences, as expected, show a greater degree of homology to the searched *Arabidopsis thaliana* sequence. NHX genes are implicated in responses to salinity in various angiosperms (Akram *et al.* 2020; Fu *et al.* 2020; Yarra 2019). The homologous sequences in the charophyte algae may have a similar function or may have arisen in those lineages and been coopted by embryophytes for their current purposes. Experimental lab work is needed to support or reject each of those hypotheses. The CHX gene family is present only in the charophyte algae, beginning with *Chlorokybus*. The *Mesostigma viride* hits to this gene family are very weak and seem to indicate that the gene family is not present, although this could be an issue of data quality. CHX genes are hypothesized to be involved with osmotic regulation and ion management during reproduction in angiosperms. It remains unknown, however, what function they might have in the different algae (Sze *et al.* 2004). The KEA gene family shares roughly the same degree of homology to the *Arabidopsis thaliana* sequence across the entire phylogeny, up to and including

the different prasinophyte taxa. KEA genes were thus present in a similar form to their current state in the ancestral Chloroplastida organism and therefore may be a core part of stress responses to ion imbalances in all green organisms (Chen *et al.* 2015; Chanroj *et al.* 2012). And, finally, the AAP searches show that the gene family is only present in embryophytes, although there is a puzzling signal in a branch of the Trebouxiophytes. Previous work done on AAP genes indicates that they are absent from the chlorophytes, although only *Volvox* and *Chlamydomonas* were examined (Tegeder and Ward 2012). This sequence may be a sign of contamination in the data, an independent evolution of a protein resembling embryophyte AAP proteins in the Trebouxiophytes, or a horizontal gene transfer event into that particular Trebouxiophyte lineage from another organism with an amino acid permease. Trebouxiophytes are soil algae and therefore may be candidates for horizontal gene transfer from soil bacteria such as *Rhizobium* (see Appendix D). These gene family searches are significant in that they represent a broad scanning of presence/absence of genes across the phylogeny, rather than relying on work focused on model organisms. Ideally, they will pave the way for future laboratory manipulation in different algae, especially those not commonly studied.

Conclusions

I have assembled and analyzed a dataset of over 60 species of Chloroplastida, including many algae species not commonly examined in the scientific literature. I successfully recover a phylogeny for these organisms and corroborate the recent placement of the Zygnematophyceae algae lineage as the sister lineage to land plants. I then show that amino acid and codon usage patterns of these species typically follow the structure of the phylogeny, although there are some notable exceptions such as the Trentepohliaes. I further demonstrate that many protein domains are shared between the charophyte algae and embryophytes, which is evidence that the

charophyte lineages were primed to colonize land. I conclude by scanning my dataset for homologous sequences to protein domains and gene families known to have important biological functions in land plants, and show that those gene families are either present through the phylogeny or arose in the early charophyte algae.

Tables

Table 1: Algae species assembled *de novo*

Species	Major Clade	Minor Clade	Media	Temperature ºC	Collection
Ankistrodesmus falcatus	Chlorophyte	Chlorophyceae	GWH	18	UTEX101
Atractomorpha echinata	Chlorophyte	Chlorophyceae	GWH	18	UTEX LB2309
Bracteacoccus aerius	Chlorophyte	Chlorophyceae	GWH	18	UTEX1250
Cephaleuros parasiticus	Chlorophyte	Ulvophyceae	Solid GWH	18	SAG 73.90
Chaetospharidium globosum	Charophyte	Coleochaetophyceae	DY III	18	SAG 26.98
Coleochaete orbicularis	Charophyte	Coleochaetophyceae	GWH	18	LB422
Elakatothrix viridis	Chlorophyte	Chlorophyceae	GWH	18	SAG 9.94
Entransia	Charophyte	Klebsormidiophyceae	DY III	18	Delwiche Lab
Eremochloris sphaerica	Chlorophyte	Trebouxiophyceae	GWH	18	Delwiche Lab
Hormotilopsis gelatinosa	Chlorophyte	Chlorophyceae	GWH	18	UTEX B104
Klebsormidium flaccidum	Charophyte	Klebsormidiophyceae	BBM	18	UTEX 321
Leptosira	Chlorophyte	Trebouxiophyceae	GWH	18	UTEX 333
Mesostigma viride	Charophyte	Mesostigmatophyceae	GWH	18	NIES 995
Mougeotia scalaris	Charophyte	Zygnematophyceae	GWH	18	SAG 164.90
Nephroselmis pyriformis	Chlorophyte	Prasinophyta	32 ppt f/2	16	CCMP 717
Nitella mirabilis lower	Charophyte	Charophyceae	N/A	N/A	N/A
Nitella_mirabilis upper	Charophyte	Charophyceae	N/A	N/A	N/A
NOT Blastophysa rhizopus	Chlorophyte	Ulvophyceae	N/A	10	KU 295
Oedogonoium cardiacum	Chlorophyte	Chlorophyceae	GWH	18	UTEX LB40
Oltmannsiellopsis unicellularis	Chlorophyte	Ulvophyceae	30 ppt L1	16	SCCAP K-2050
Oocystis solitaria	Chlorophyte	Trebouxiophyceae	GWH	18	SAH 83.80
Penium margaritaceum	Charophyte	Zygnematophyceae	GWH	18	SAG 22.82
Phaeophila dendroides	Chlorophyte	Ulvophyceae	32 ppt L1	18	CCMP 2372
Prasiolopsis	Chlorophyte	Trebouxiophyceae	Solid GWH	18	SAG 84-81
Pyramimonas parkeae	Chlorophyte	Prasinophyta	32 ppt f/2	16	CCMP 726
Spirogyra pratensis	Charophyte	Zygnematophyceae	GWH	18	UTEX 921
Spirogyra Au1	Charophyte	Zygnematophyceae	GWH	18	Delwiche Lab
Tetraselmis striata	Chlorophyte	Chlorodendrophyceae	32 ppt f/2	18	SAG 41.85
Tetraselmis suecica	Chlorophyte	Chlorodendrophyceae	32 ppt f/2	18	PLY 305
Trebouxia aggregata	Chlorophyte	Trebouxiophyceae	GWH	18	SAG 219-1d
Trentepohlia annulata	Chlorophyte	Ulvophyceae	Solid GWH	18	SAG 20.94
Watanabea reniformis	Chlorophyte	Trebouxiophyceae	GWH	18	SAG 211-9b

Table 2: Additional taxa included in the phylogenetic analysis

Species	Major Clade	Minor Clade	Data Type	Data Source
Amborella trichopoda	Embryophyte	Angiosperm	Genome	NCBI
Arabidopsis thaliana	Embryophyte	Angiosperm	Genome	NCBI
Azolla filiculoides	Embryophyte	Angiosperm	Genome	Fernbase
Blasia	Embryophyte	Pre-Vascular	Transcriptome	Collaborator
Chara australis	Charophyte	Charophyceae	Transcriptome	Collaborator
Chara braunii	Charophyte	Charophyceae	Genome	NCBI
Chara longifolia	Charophyte	Charophyceae	Transcriptome	Collaborator
Chlorokybus atmophyticus	Charophyte	Chlorokybophyceae	Transcriptome	OneKP
Codium fragile	Chlorophyte	Ulvophyceae	Transcriptome	OneKP
Coleochaete irregularis	Charophyte	Coleochaetophyceae	Transcriptome	OneKP
Cycas micholitzii	Embryophyte	Gymnosperm	Transcriptome	OneKP
Ginkgo biloba	Embryophyte	Gymnosperm	Transcriptome	OneKP
Gnetum montanum	Embryophyte	Gymnosperm	Transcriptome	OneKP
Marchantia polymorpha	Embryophyte	Pre-Vascular	Genome	Genome Website
Marsilea	Embryophyte	Early Vascular	Transcriptome	Collaborator
Mesotaenium endlicherianum	Charophyte	Zygnematophyceae	Genome	Genome Website
Nothoceros aenigmaticus	Embryophyte	Pre-Vascular	Transcriptome	OneKP
Nothoceros vincentianus	Embryophyte	Pre-Vascular	Transcriptome	OneKP
Oryza sativa	Embryophyte	Angiosperm	Genome	NCBI
Ostreococcus lucimarinus	Chlorophyte	Prasinophyta	Genome	NCBI
Ostreococcus tauri	Chlorophyte	Prasinophyta	Genome	NCBI
Penium margaritaceum	Charophyte	Zygnematophyceae	Genome	Genome Website
Physcomitrella patens	Embryophyte	Pre-Vascular	Genome	NCBI
Pinus taeda	Embryophyte	Gymnosperm	Transcriptome	OneKP
Salvinia cucullata	Embryophyte	Early Vascular	Genome	Fernbase
Selaginella moellendorffii	Embryophyte	Early Vascular	Genome	NCBI
Sorghum bicolor	Embryophyte	Angiosperm	Genome	NCBI
Spiroglea muscilosa	Charophyte	Zygnematophyceae	Genome	Genome Website
Taxus baccata	Embryophyte	Gymnosperm	Transcriptome	OneKP
Welwitschia mirabilis	Embryophyte	Gymnosperm	Transcriptome	OneKP
Zea mays	Embryophyte	Angiosperm	Genome	NCBI

Table 3: Algal genomes used in hidden Markov model construction for ortholog identification

Species	Major Clade	Minor Clade	Data Source
Chara braunii	Charophyte	Charophyceae	NCBI
Chlamydomonas reinhardtii	Chlorophyte	Chlorophyceae	NCBI
Dunaliella salina	Chlorophyte	Chlorophyceae	NCBI
Gonium pectorale	Chlorophyte	Chlorophyceae	NCBI
Klebsormidium nitens	Charophyte	Klebsormidiophyceae	NCBI
Monoraphidium neglectum	Chlorophyte	Chlorophyceae	NCBI
Penium margaritaceum	Charophyte	Zygnematophyceae	Genome Website
Raphidocelis subcapitata	Chlorophyte	Chlorophyceae	NCBI
Volvox carteri	Chlorophyte	Chlorophyceae	NCBI

Grouping	Taxa
Charophyceae s. s.	*Chara australis, Chara braunii, Chara longifolia, Nitella mirabilis*
Coleochaetophyceae	*Chaetosphaeridium globosum, Coleochaete irregularis, Coleochaete orbicularis*
Embryophytes	*Amborella trichopoda, Arabidopsis thaliana, Azolla filiculoides, Blasia, Cycas micholitzii, Ginkgo biloba, Gnetum montanum, Marchantia polymorpha, Marsilea, Nothoceros aenigmaticus, Nothoceros vincentianus, Oryza sativa, Physcomitrella patens, Pinus taeda, Salvinia cucullata, Selaginella moellendorffii, Sorghum bicolor, Taxus baccata, Welwitschia mirabilis, Zea mays*
Older Taxa	*Ankistrodesmus falcatus, Atractomorpha echinata, Bracteacoccus aerius, Cephaleuros parasiticus, Chlorokybus atmophyticus ,Codium fragile, Elakatothrix viridis, Entransia Eremochloris sphaerica, Hormotilopsis gelatinosa, Klebsormidium flaccidum, Leptosira, Mesostigma viride, NOT Blastophysa rhizopus, Nephroselmis pyriformis, Oedogonium cardiacum, Oltmannsiellopsis unicellularis, Oocystis solitaria, Ostreococcus lucimarinus, Ostreococcus tauri Phaeophila dendroides, Prasiolopsis, Pyramimonas parkeae, Tetraselmis striata, Tetraselmis suecica, Trebouxia aggregata, Trentepohlia annulata, Watanabea reniformis*
Zygnematophyceae	*Mesotaenium endlicherianum, Mougeotia scalaris, Penium margaritaceum genome, Penium margaritaceum, Spirogloea muscicola, Spirogyra Aus, Spirogyra pratensis*

Figures

Figure 1: BUSCO Scores of Taxa in the Phylogenetic Analysis

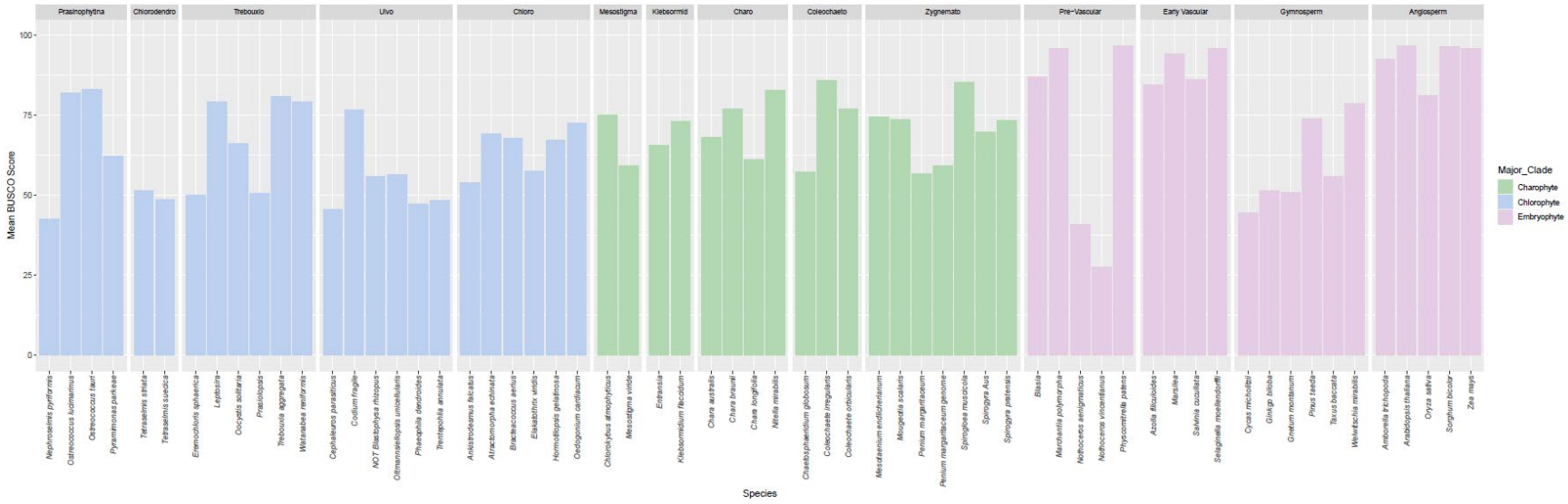


Figure 1: Mean BUSCO scores across the Viridiplantae, Chlorophyte, and Eukaryote BUSCO databases for each taxa in the phylogenetic dataset. Clade abbreviations going from left to right: Chlorodendro=Chlorodendrophyceae, Trebouxio=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, Mesostigma=Mesostigmatophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 2: BUSCO Scores for Algal Genomes

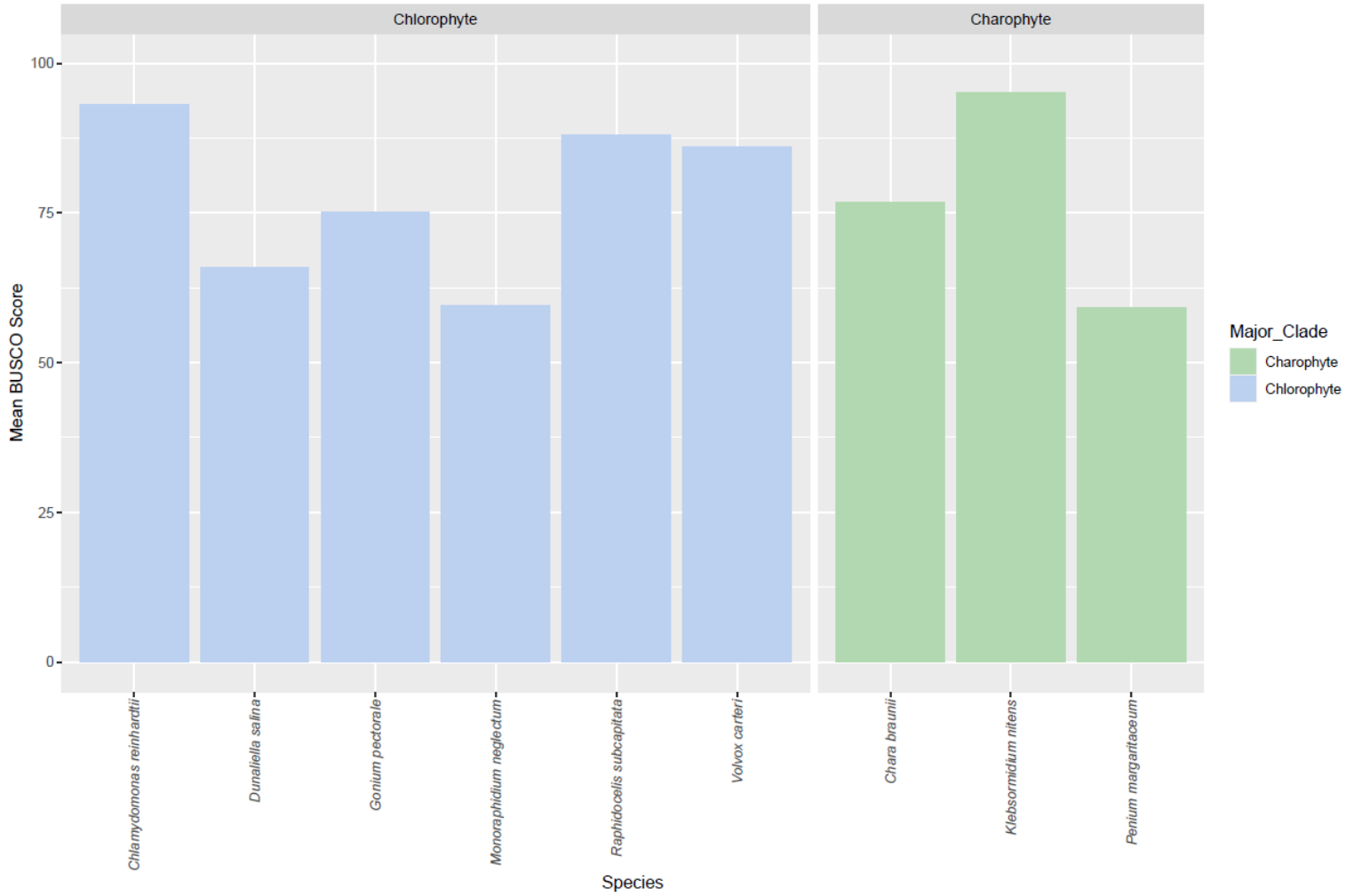


Figure 2: Mean BUSCO scores across the Viridiplantae, Chlorophyte, and Eukaryote BUSCO databases for each algal taxa used for orthogroup identificaiton and hidden Markov model construction.

Figure 3: Orthogroup Presence Across Taxa

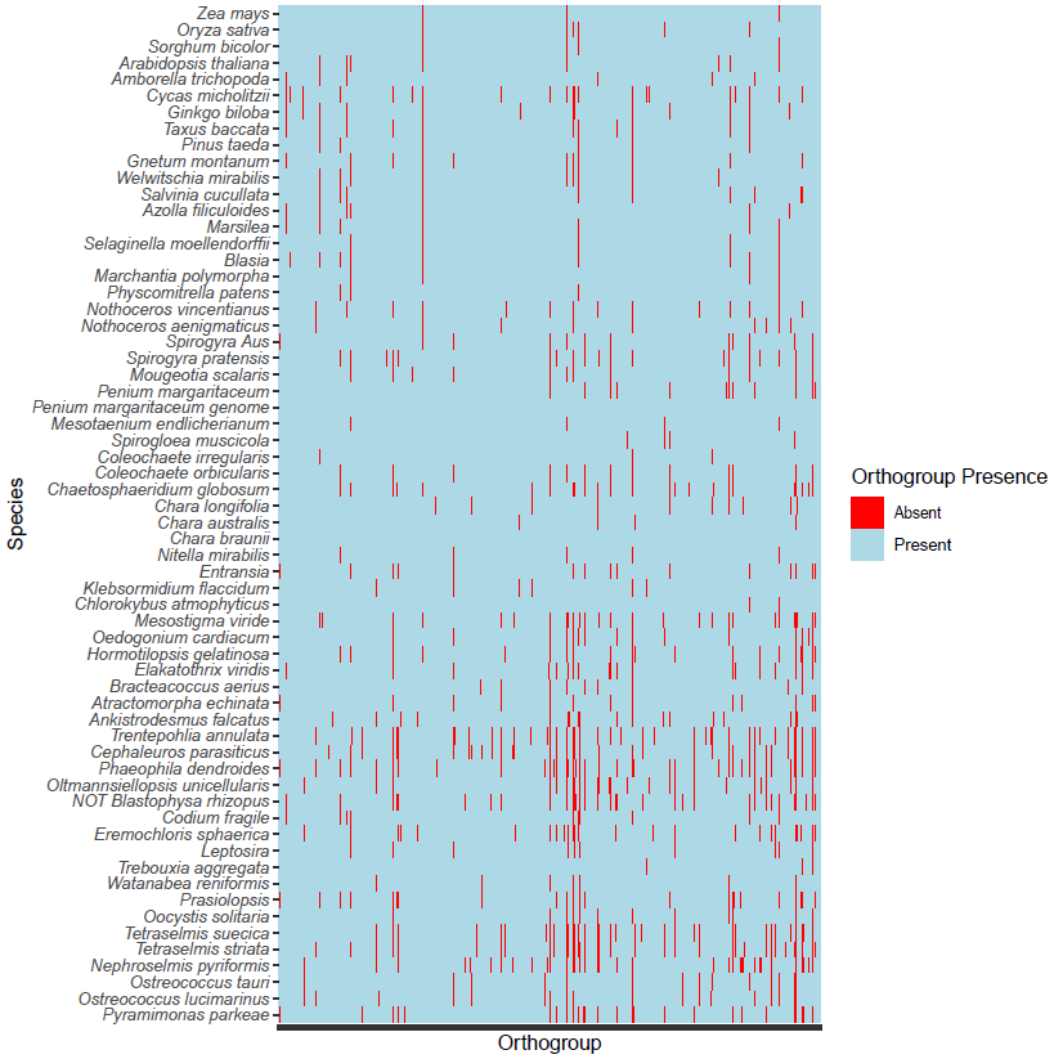


Figure 3: Orthogroup presence and absence across taxa in the phylogenetic dataset. The *Penium margaritaceum* and *Chara braunii* genomes were used in both HMM construction and in the later phylogenetic analysis.

Figure 4: Percentage of Gaps in the SuperAlignment

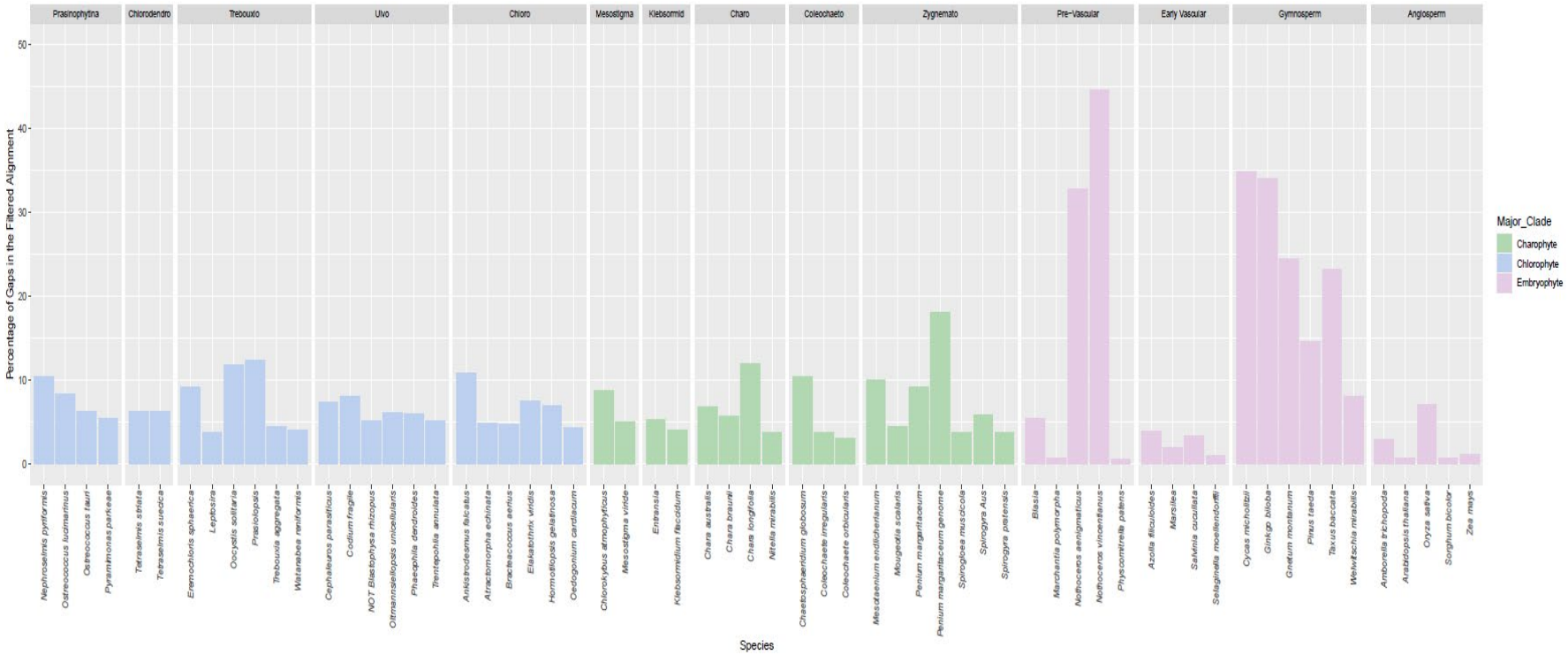


Figure 4: Percentage of gaps in each taxa in the phylogenetic analysis after trimAl filtering.

Figure 5: Phylogenetic Tree from the LG+F+R10 Analysis

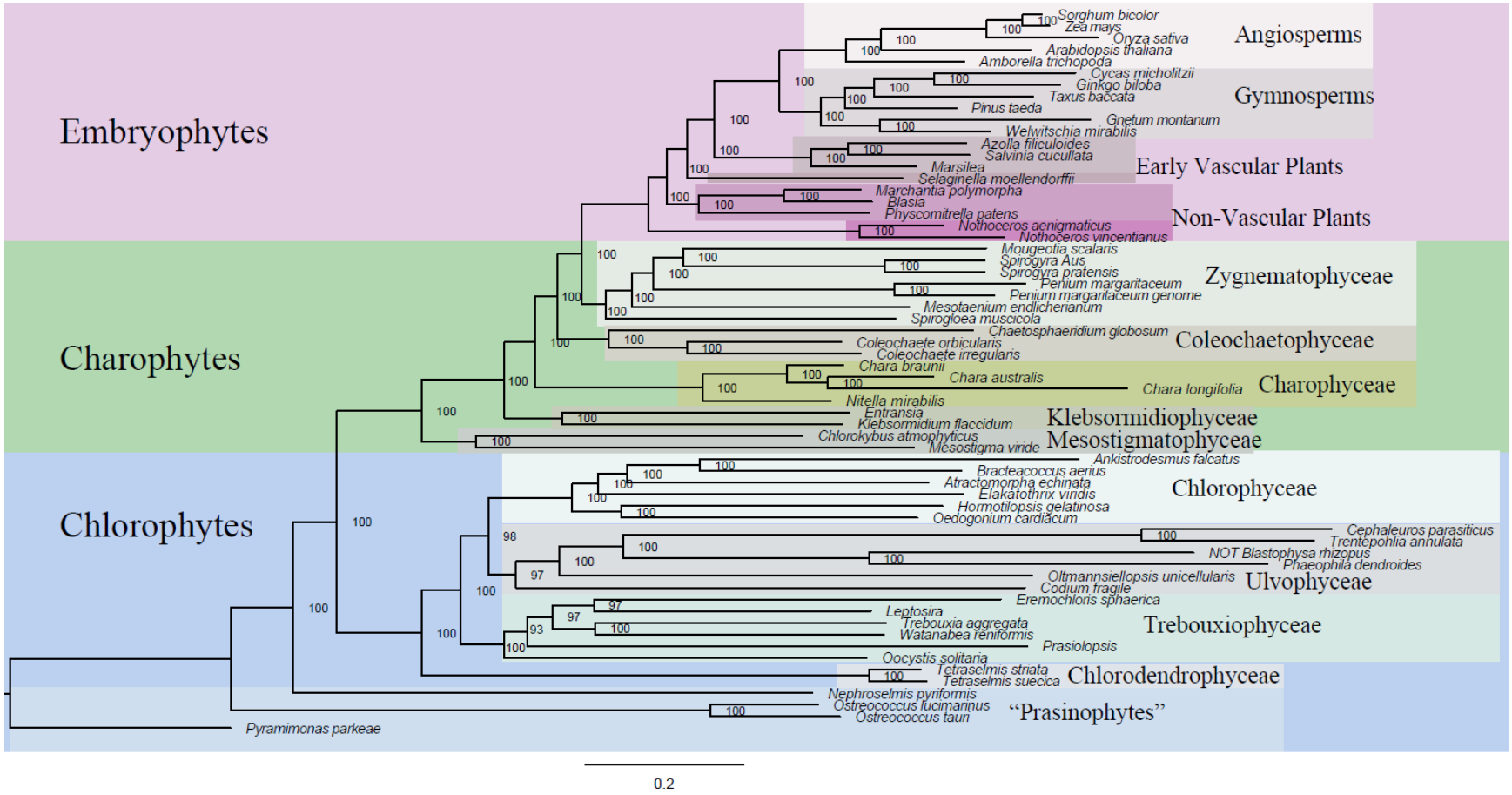


Figure 5: Phylogenetic tree from the IQ-TREE LG+F+R10 analysis. The tree was rooted at *Pyramimonas parkeae*. Bootstrap support values from ten thousand ultrafast bootstrap replicates are shown at each node.

Figure 6: Phylogenetic Tree from the LG+F*H4 Analysis

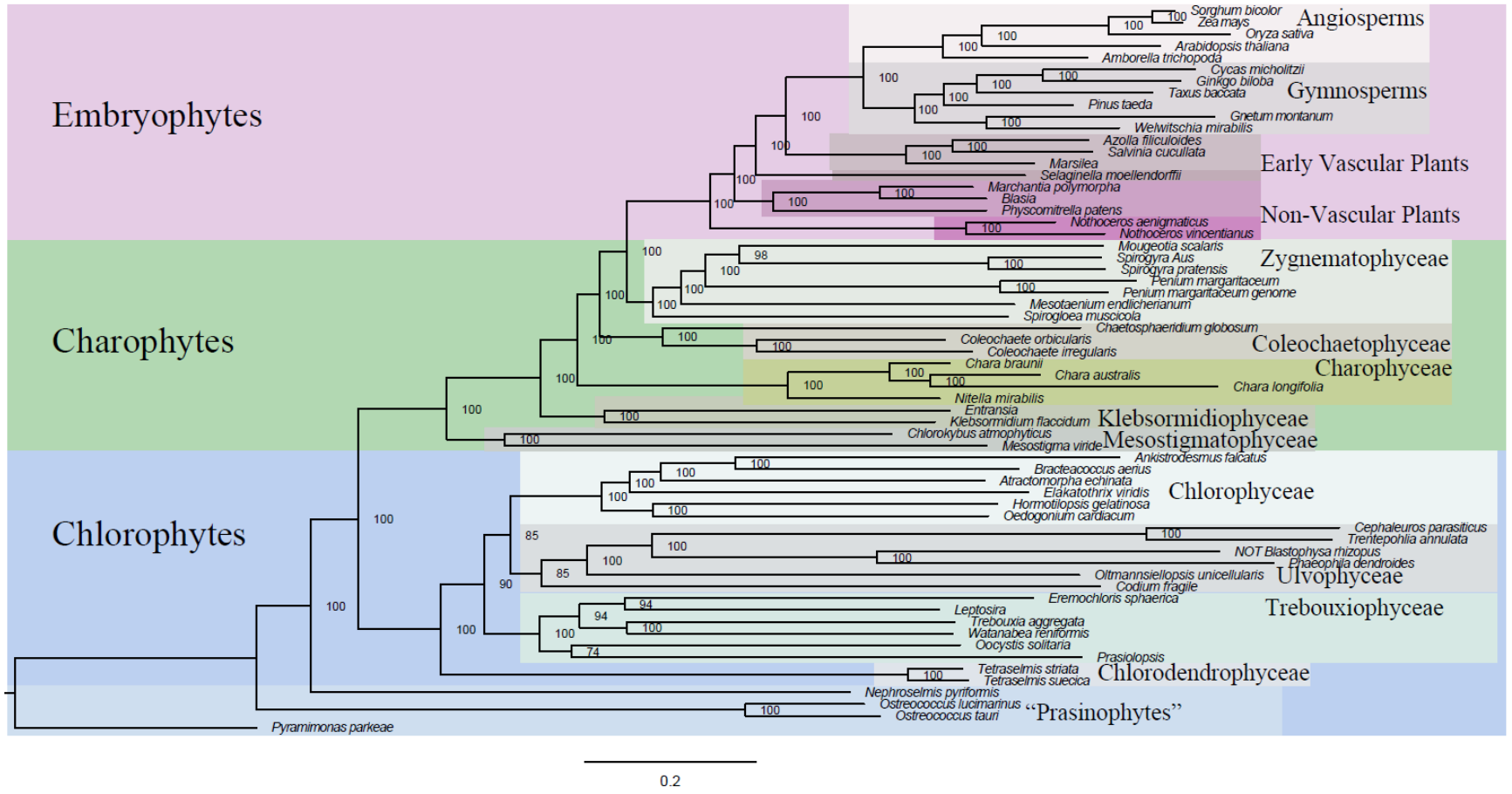


Figure 6: Phylogenetic tree from the IQ-TREE LG+F*H4 analysis. The tree was rooted at *Pyramimonas parkeae*. Bootstrap support values from ten thousand ultrafast bootstrap replicates are shown at each node.

Figure 7: Distribution of Optimal Models for each individual Gene Tree

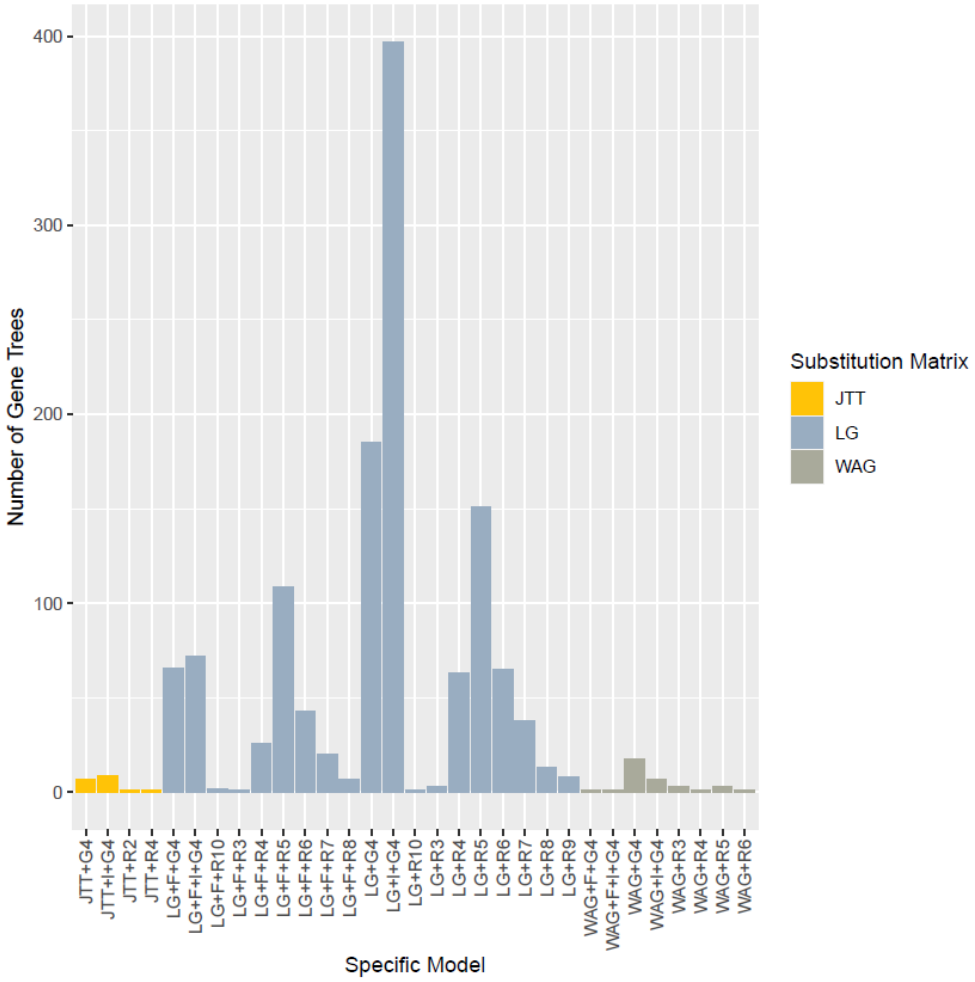


Figure 7: Distribution of gene models found for each of the 1323 gene trees. Gene trees are organized by the amino acid substitution matrix (JTT, LG, or WAG) and number and type of rate heterogeneity parameters. I indicates that the model uses a proportion of invariant sites, F indicates that the model uses empirical amino acid frequencies, G indicates that the rate parameters are drawn from a gamma distribution, and R indicates that the distribution of rate parameters is nonparametric.

Figure 8: Phylogenetic Tree from the Partitioned Analysis

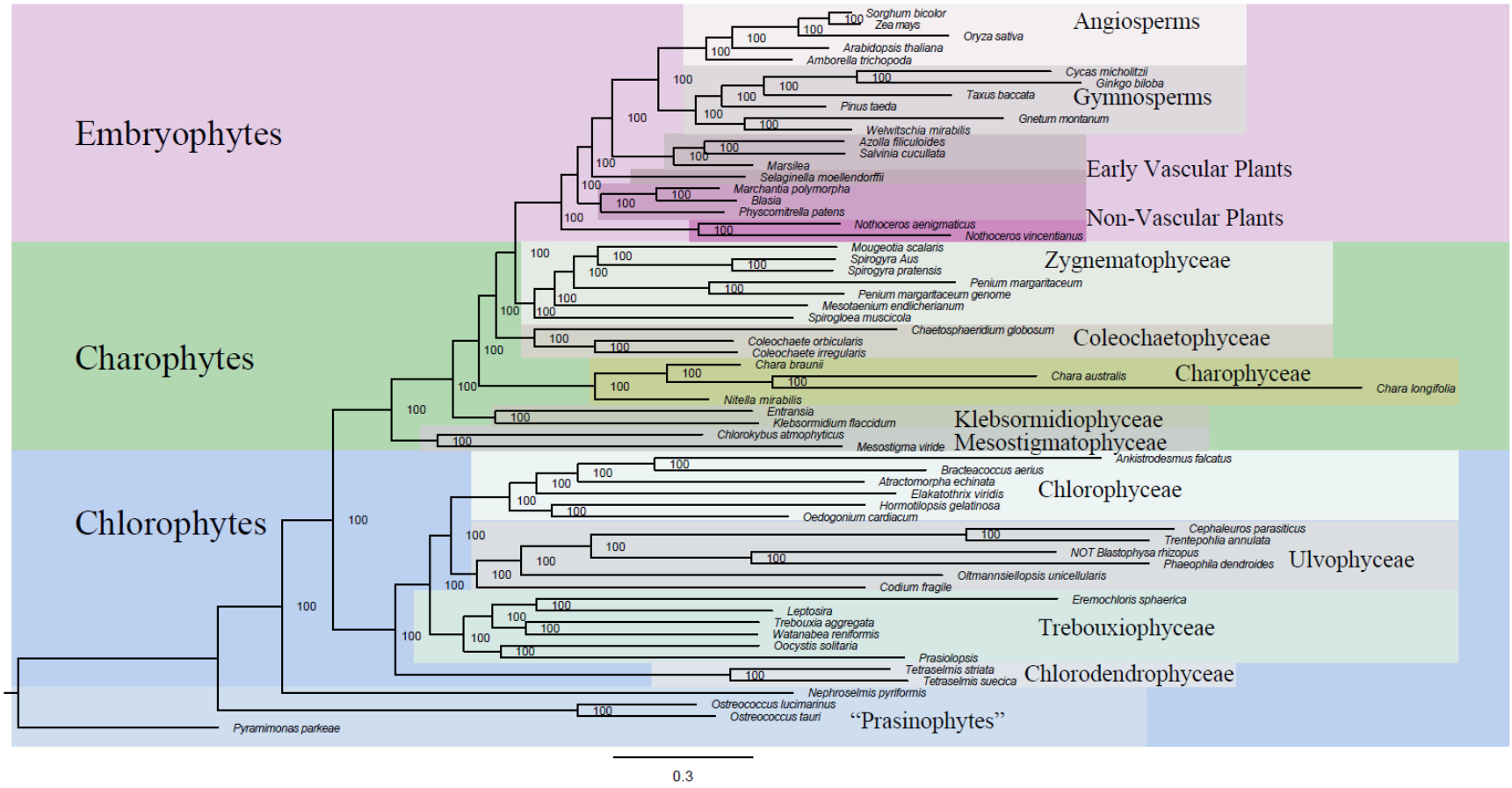


Figure 8: Phylogenetic tree from the IQ-TREE partitioned analysis. The tree was rooted at *Pyramimonas parkeae*. Bootstrap support values from ten thousand ultrafast bootstrap replicates are shown at each node.

Figure 9: ASTRAL Phylogenetic Tree

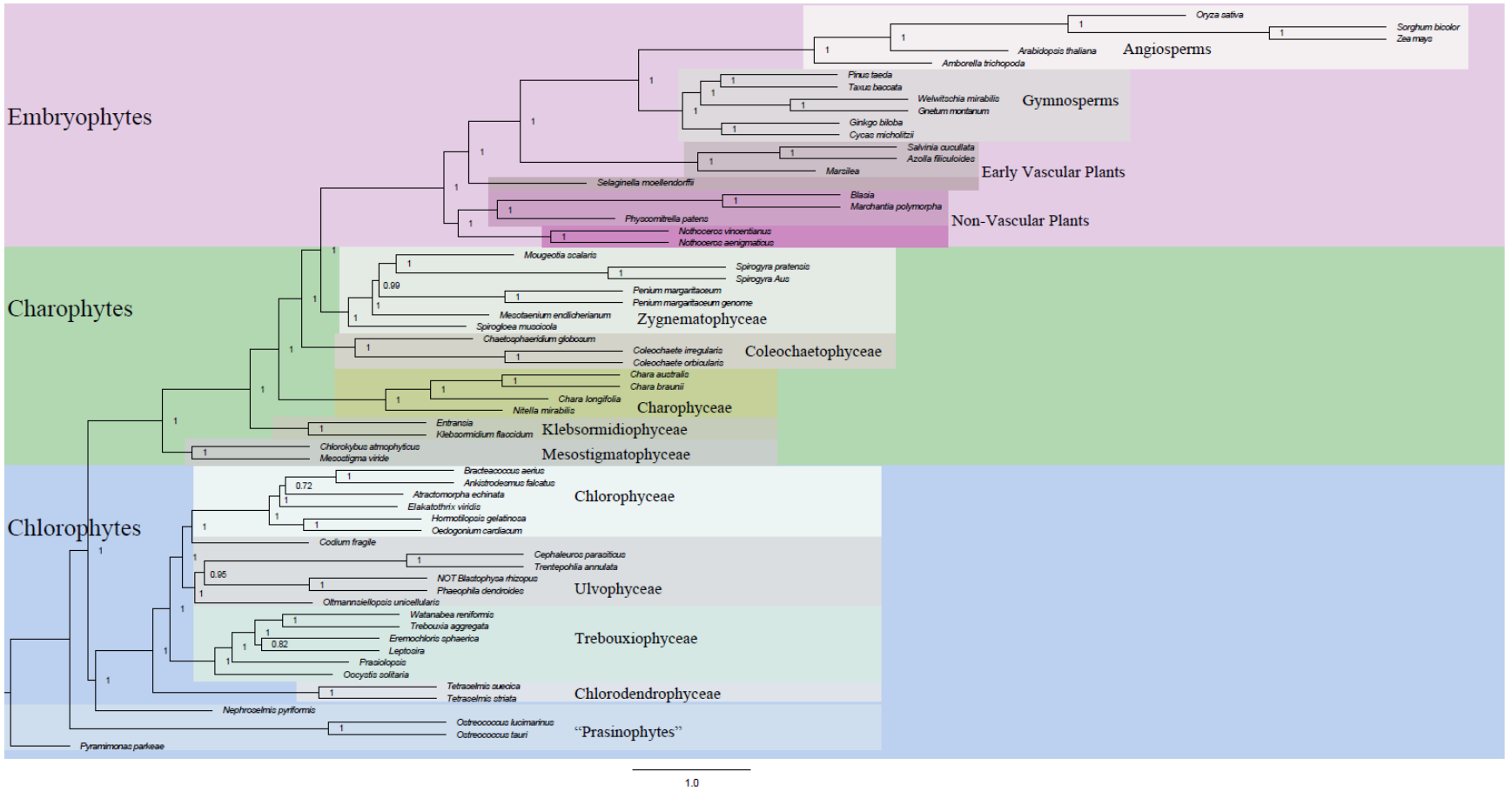


Figure 9: ASTRAL phylogenetic tree constructed with the 1323 individual gene trees. The tree was rooted at *Pyramimonas parkeae*. Node labels show the local posterior probability for that branch.

Figure 10: Likelihood Mapping Analysis

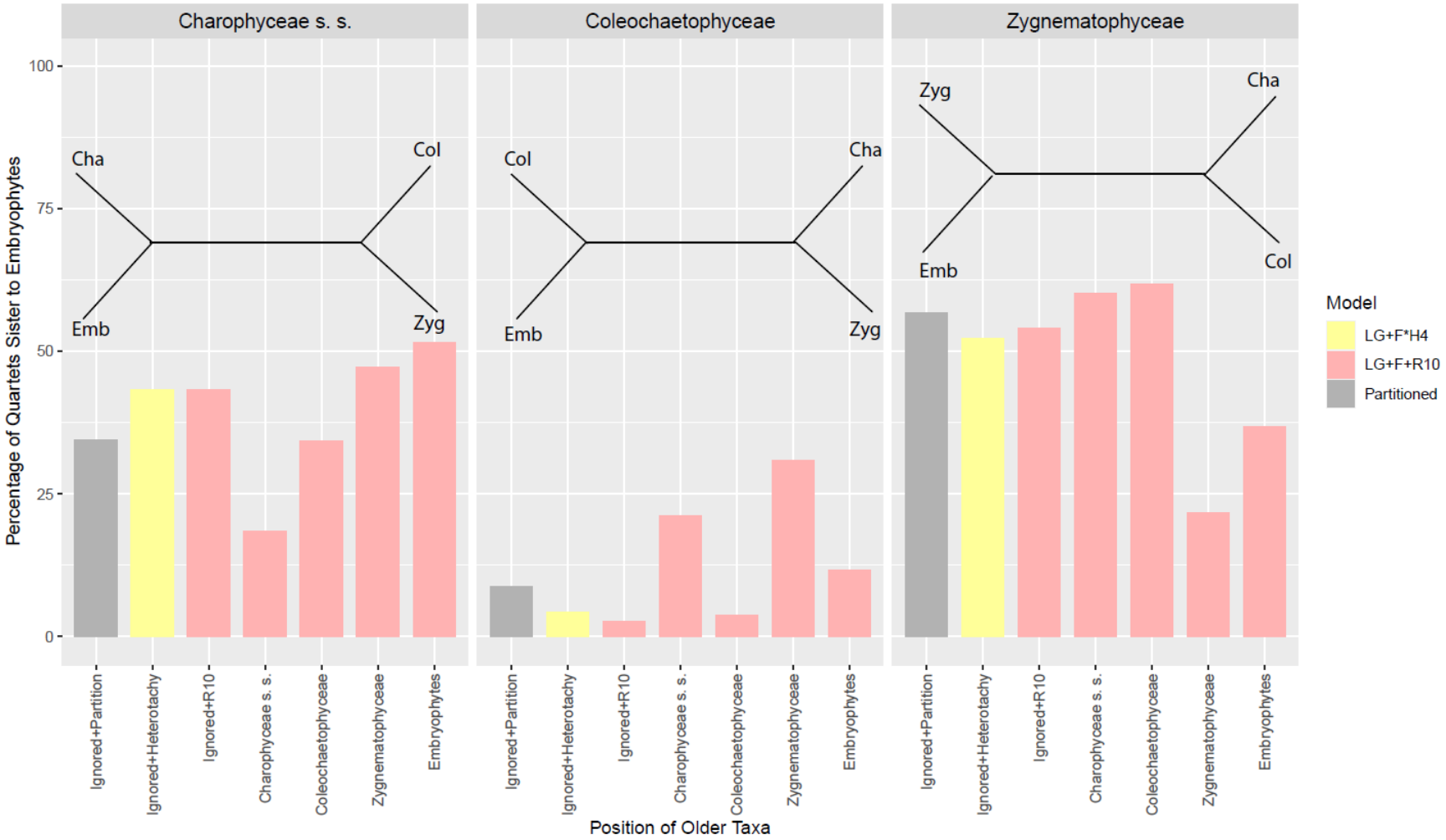


Figure 10: Percentage of likelihood quartets that support each of three charophyte lineages as the sister lineage to embryophytes. The corresponding four-taxon tree topology is shown at the top of each panel. The x axis describes where older taxa are included in the likelihood mapping analysis. The number of likelihood quartets for each different grouping of Older Taxa is as follows: Ignored = 1680, Charophyceae s. s. = 13440, Coleochaetophyceae = 17360, Zygnematophyceae = 8400, Embryophytes = 4032.

Figure 11: Analysis of 15 Possible Five-Taxon Trees

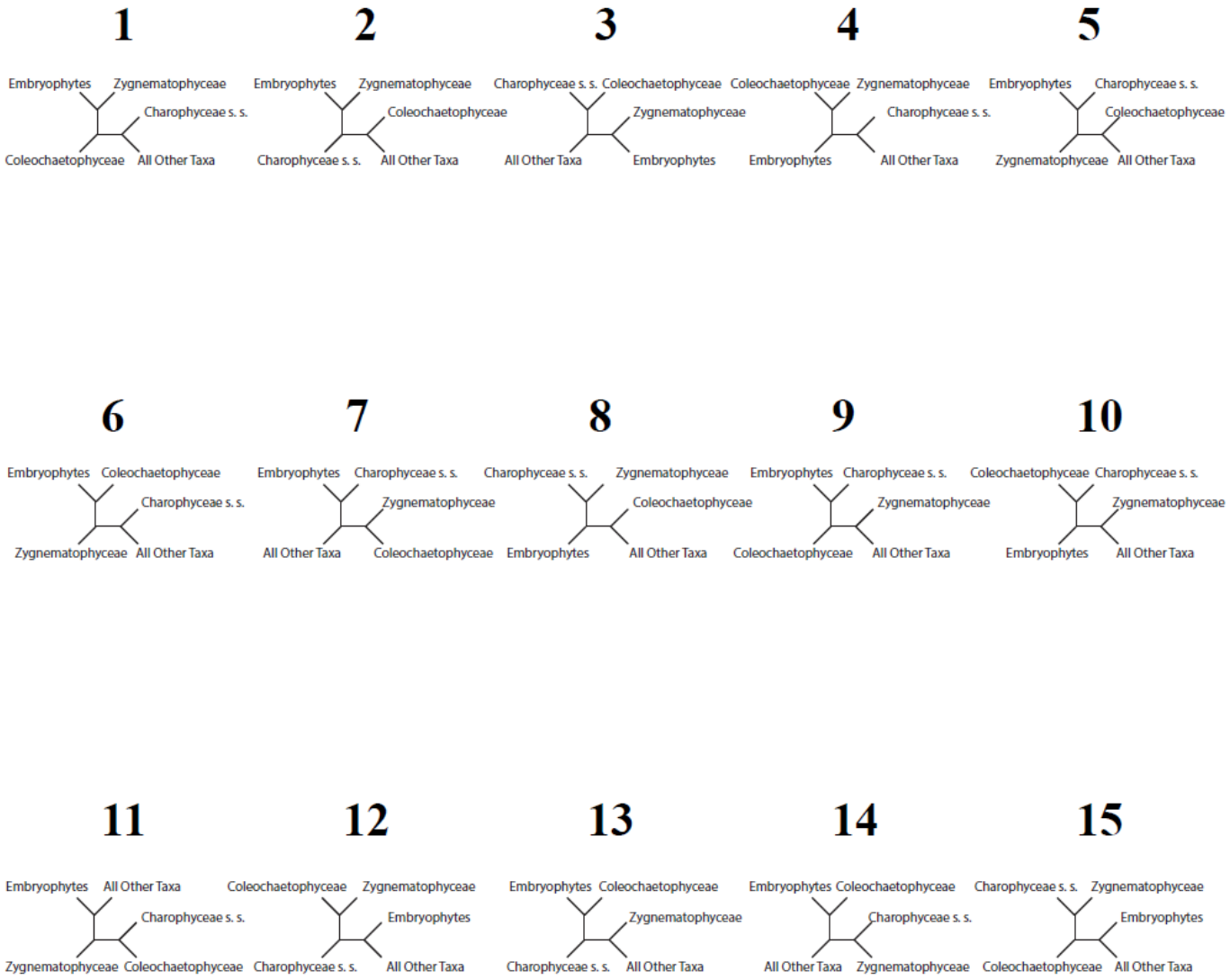
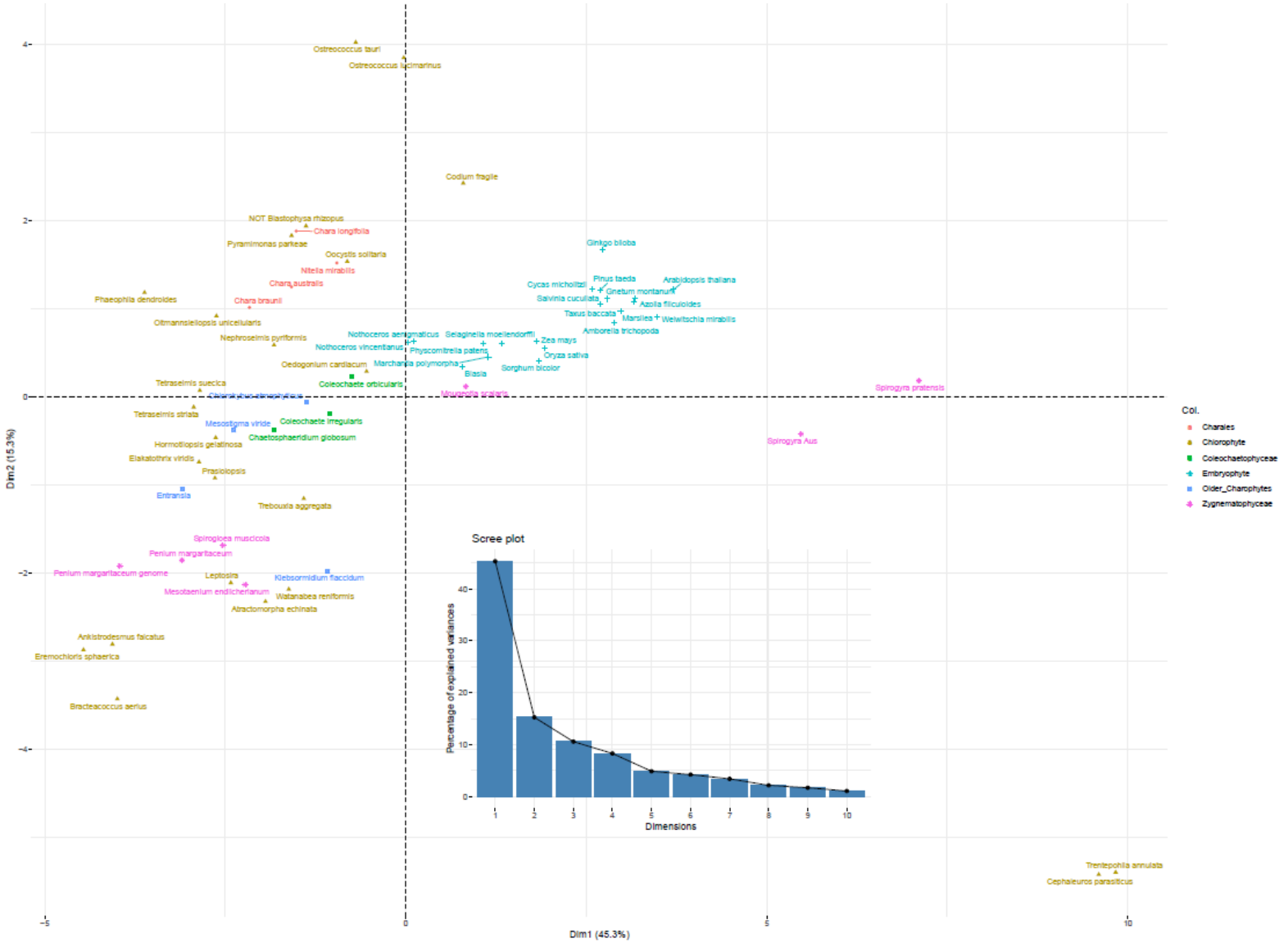


Figure 11: Ranking, according to log-likelihood, of each of the 15 possible 5-taxon arrangements of the Charophyceae s. s., Coleochaetophyceae, Zygnematophyceae, Embryophyte, and Older Taxa groupings. Trees are ranked in descending order.

Figure 12: Principal Components Analysis of Amino Acid Frequencies in Orthologous Genes



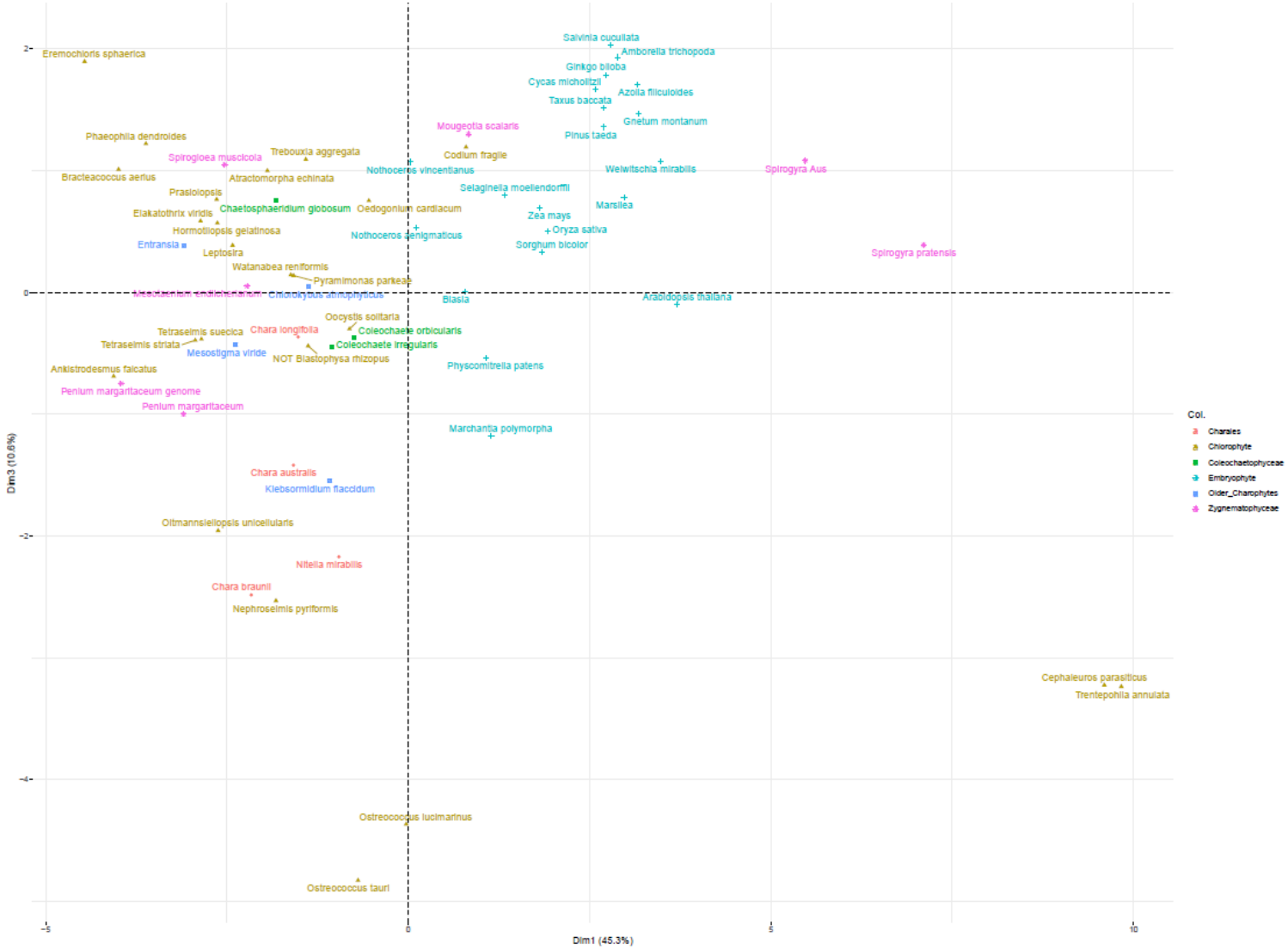
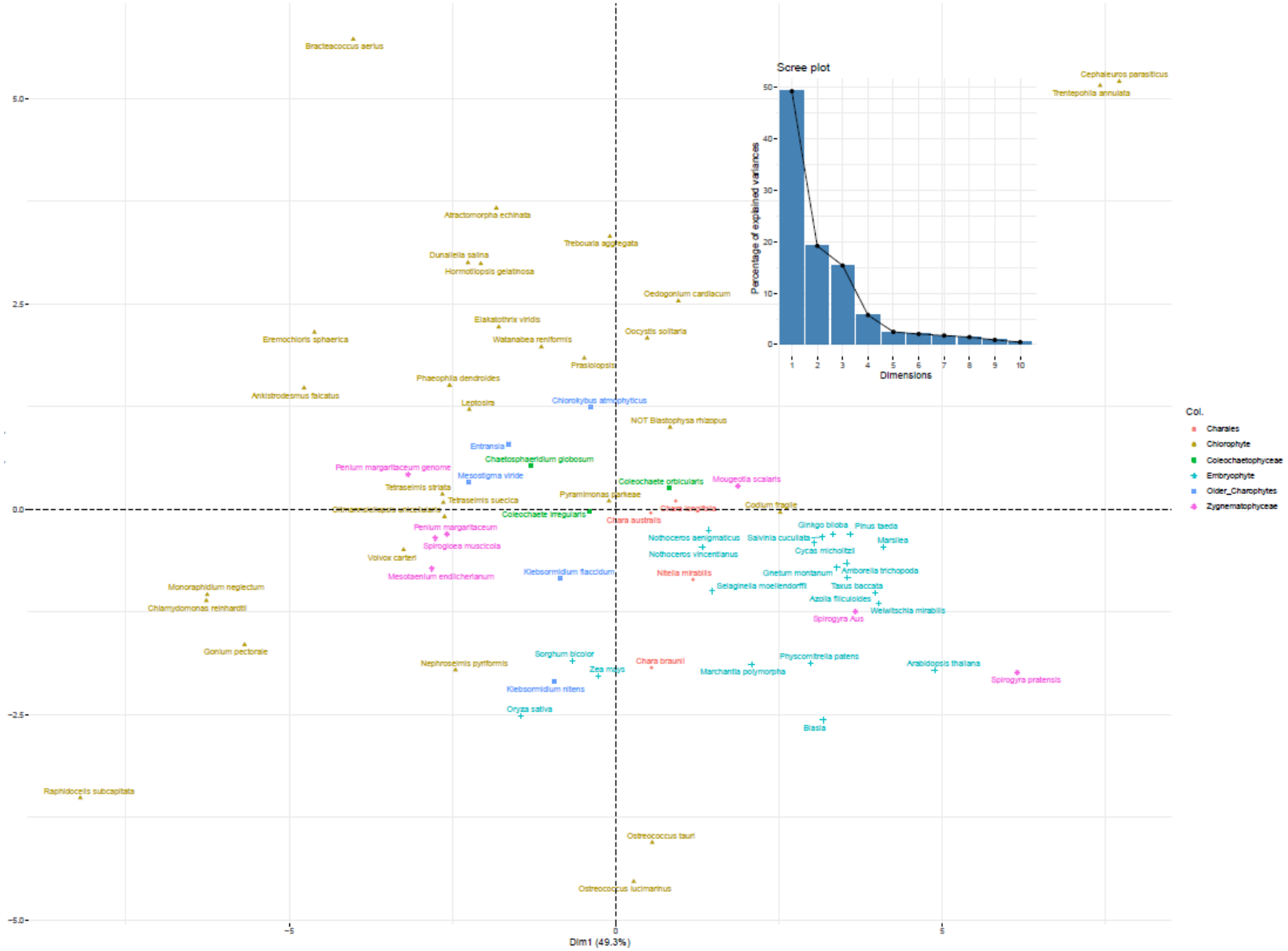


Figure 12: Dimensions 1, 2, and 3 of the principal components analysis of orthologous genes included in the multiple sequence alignment. Inset scree plot shows the percentage of variation explained by each dimension.

Figure 13: Principal Components Analysis of Amino Acid Frequencies in Whole Assemblies



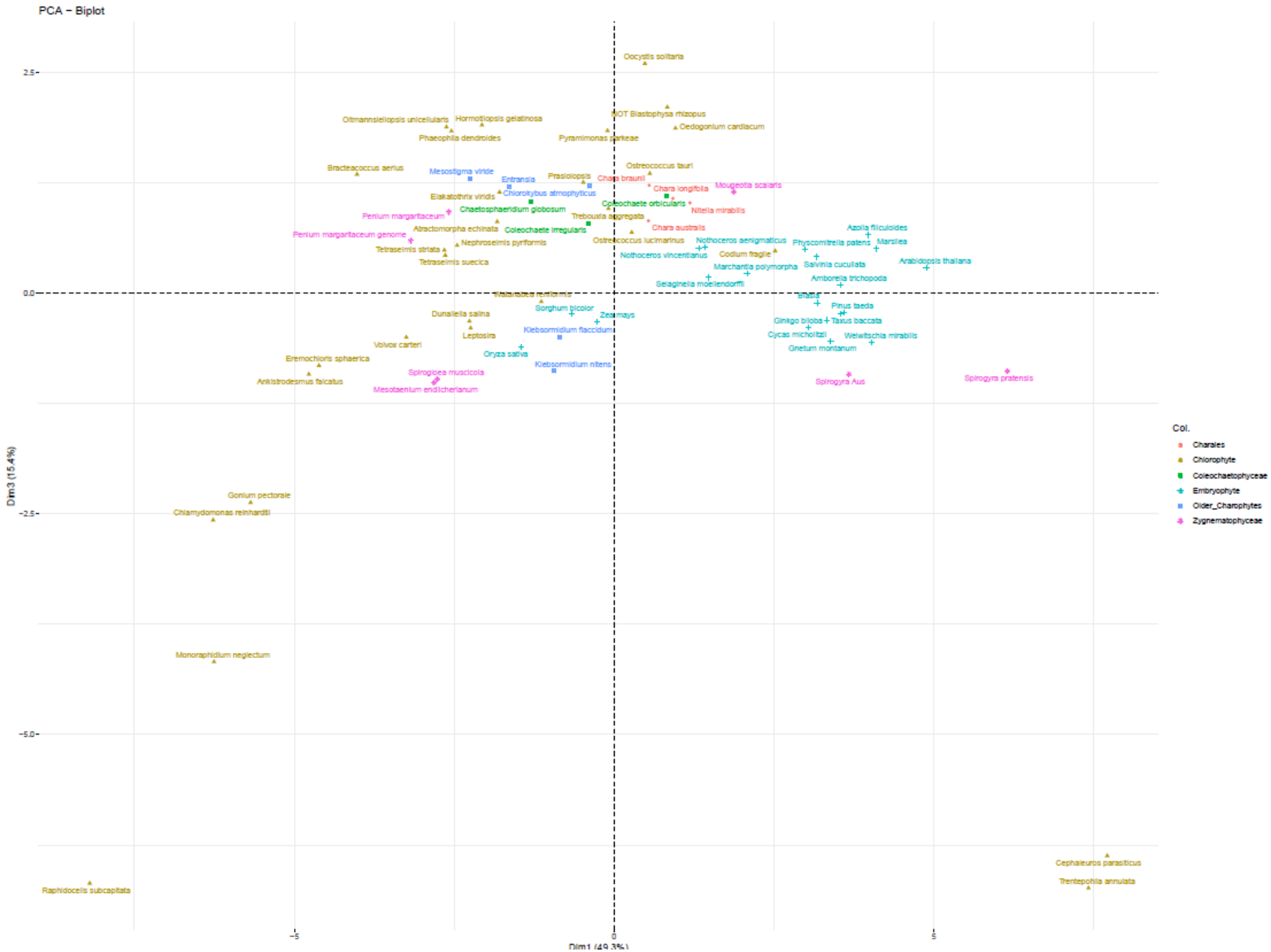
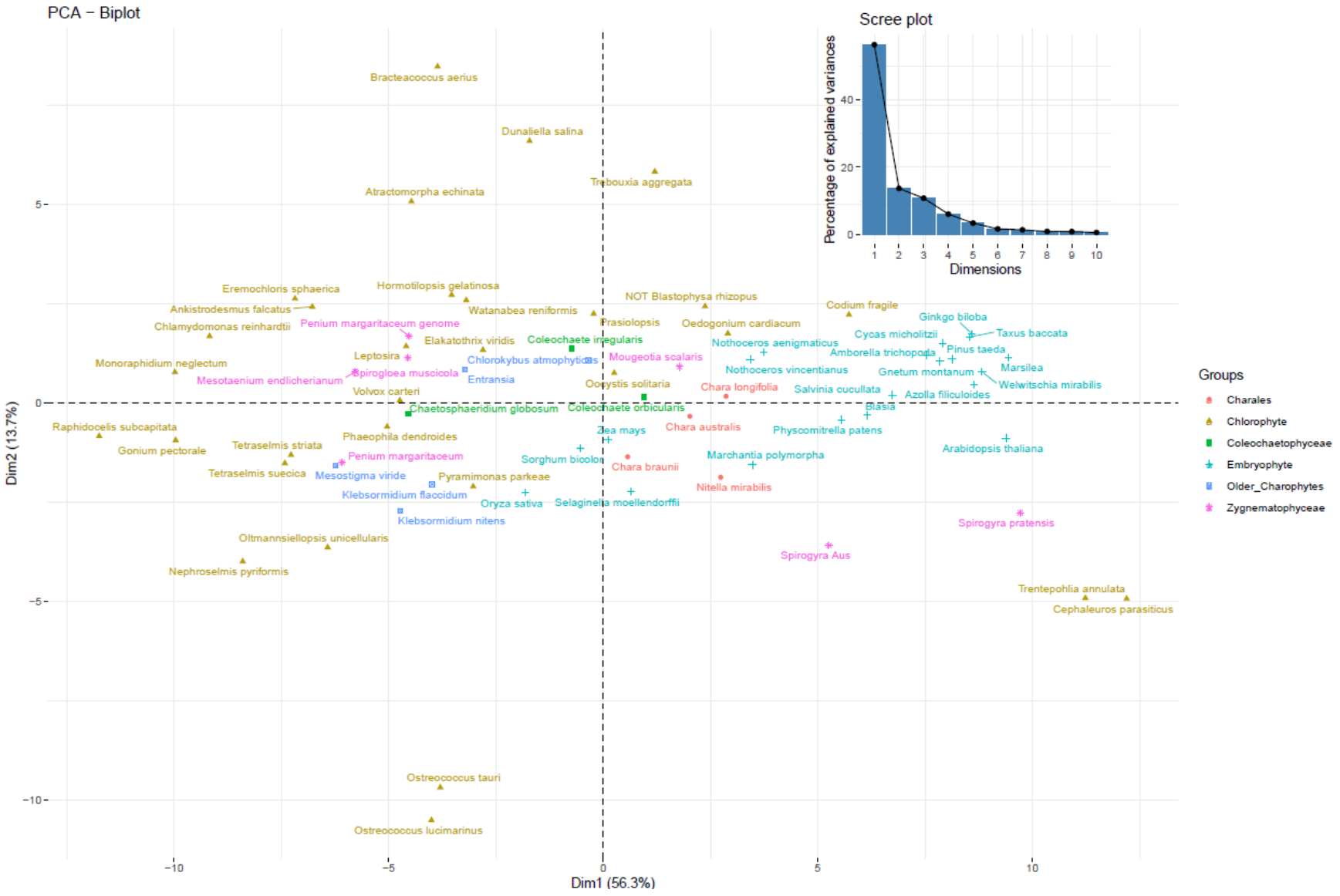


Figure 13: Dimensions 1, 2, and 3 of the principal components analysis of amino acid frequencies in the genomic and transcriptomic datasets. Inset scree plot shows the percentage of variation explained by each dimension.

Figure 14: Principal Components Analysis of Codon Frequencies in Whole Assemblies



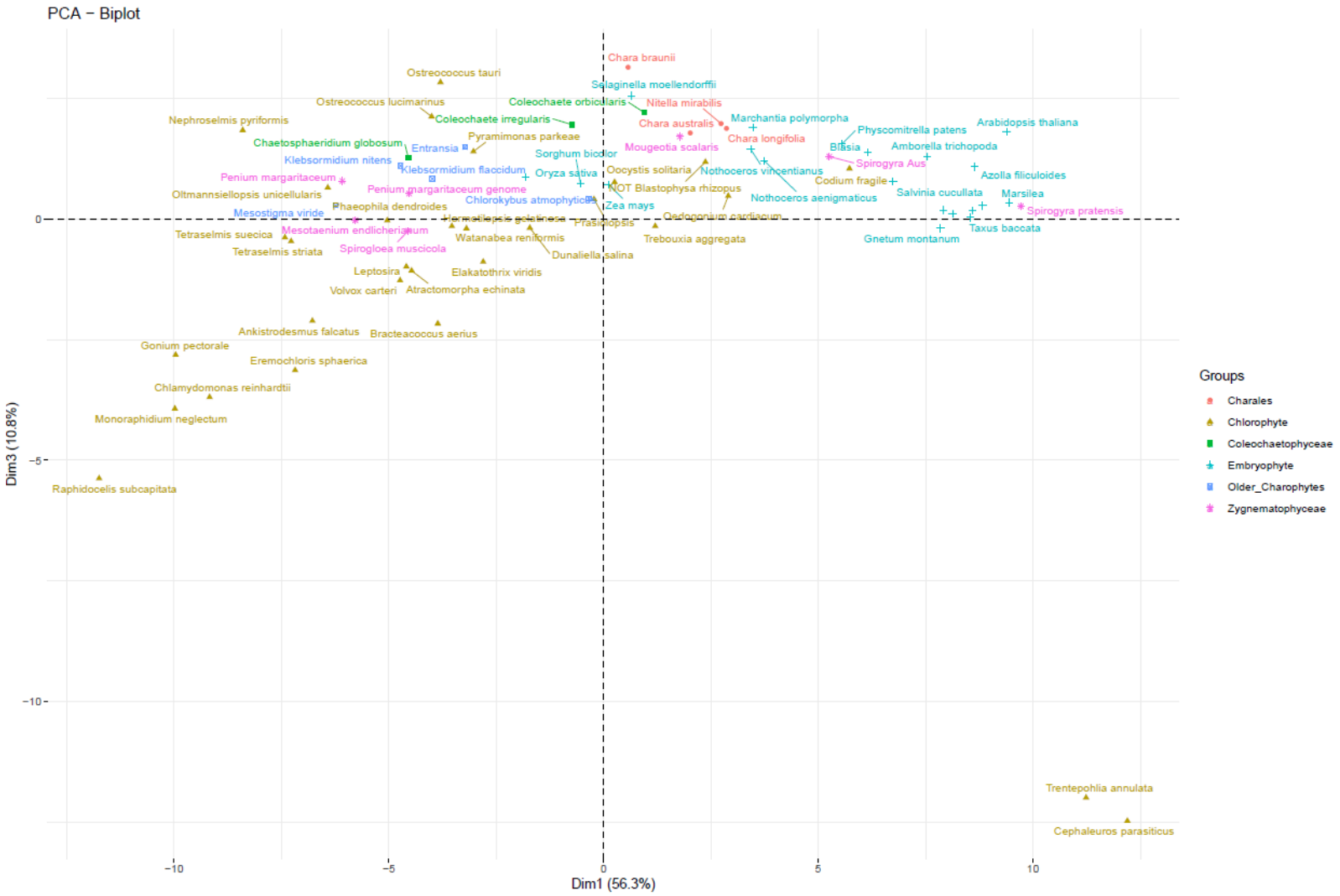


Figure 14: Dimensions 1, 2, and 3 of the principal components analysis of codon frequencies in the genomic and transcriptomic datasets. Inset scree plot shows the percentage of variation explained by each dimension.

Figure 15: Upset Plot of Protein Domain Searches

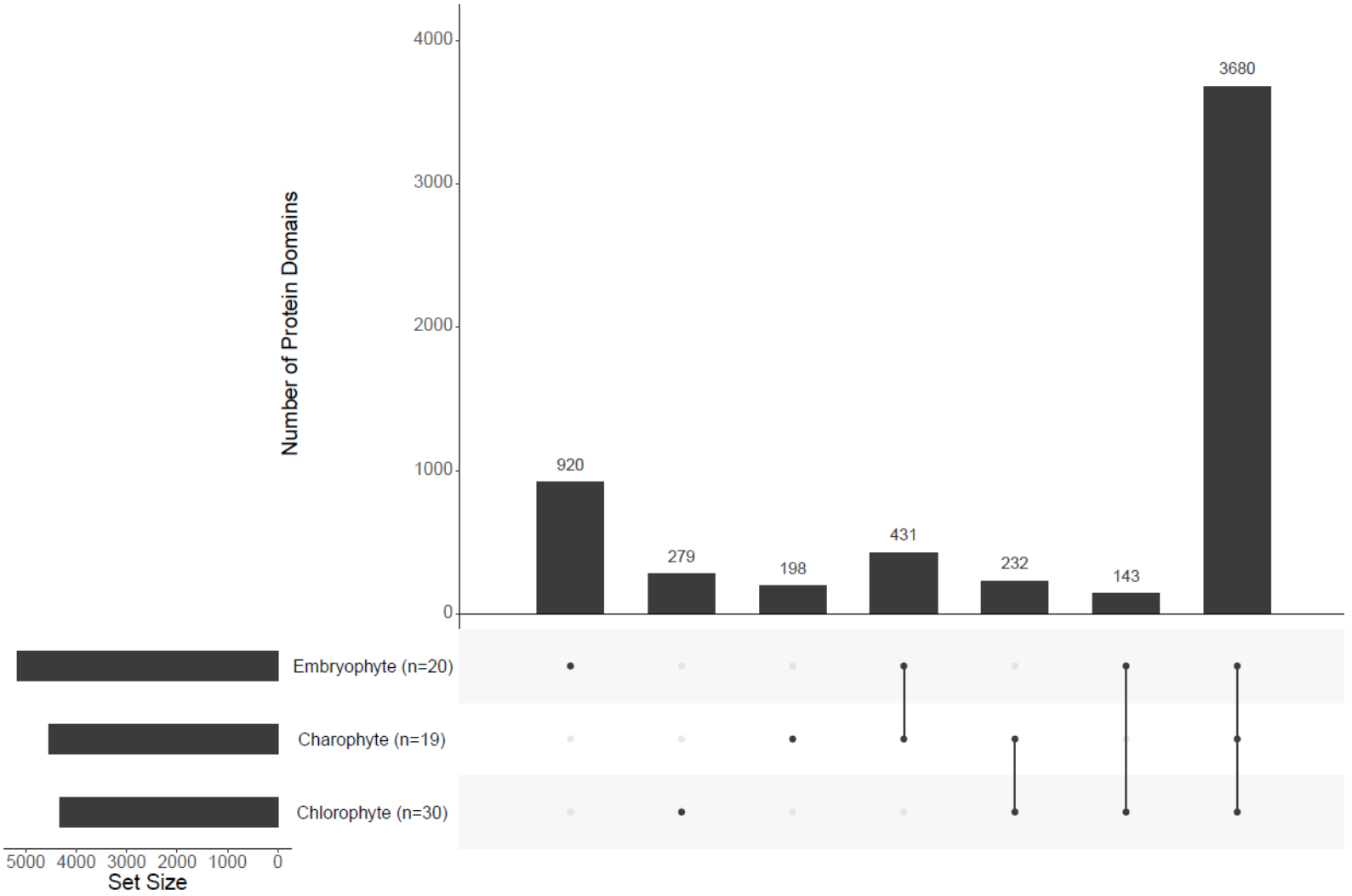


Figure 15: Upset plot of protein domain searches from the conserved domain database. The number of taxa belonging to each of the major clades is included on the bottom left.

Figure 16: CTR1 N-terminal Domain Search

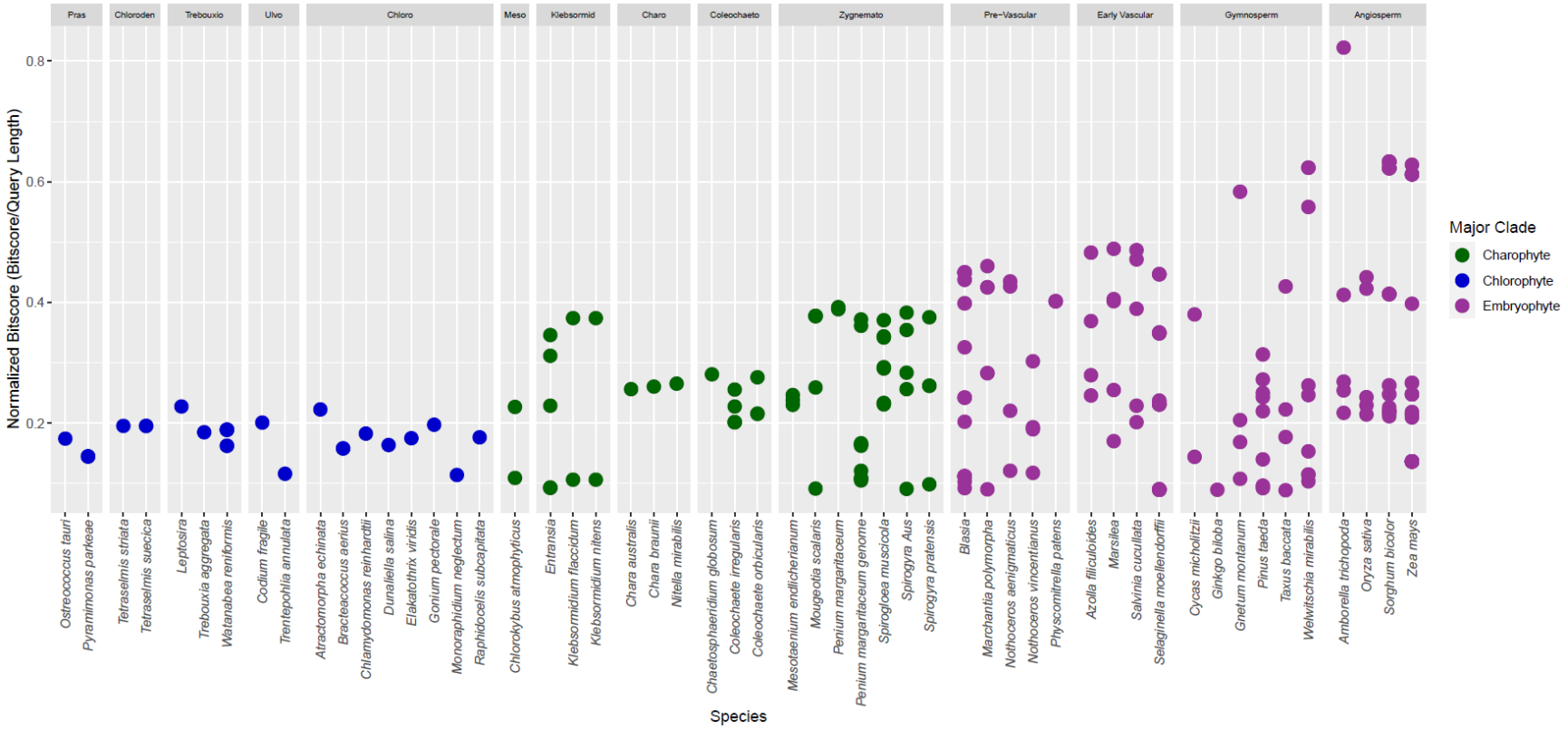


Figure 16: BLAST search of the dataset with the CTR1 N-terminal domain of the *Arabidopsis thaliana* sequence. Clade abbreviations going from left to right: Pras=Prasinophytes, Chloroden=Chlorodendrophyceae, Trebouxio=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, Meso=Mesostigmatophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 17: EIN2 Signaling Domain Search

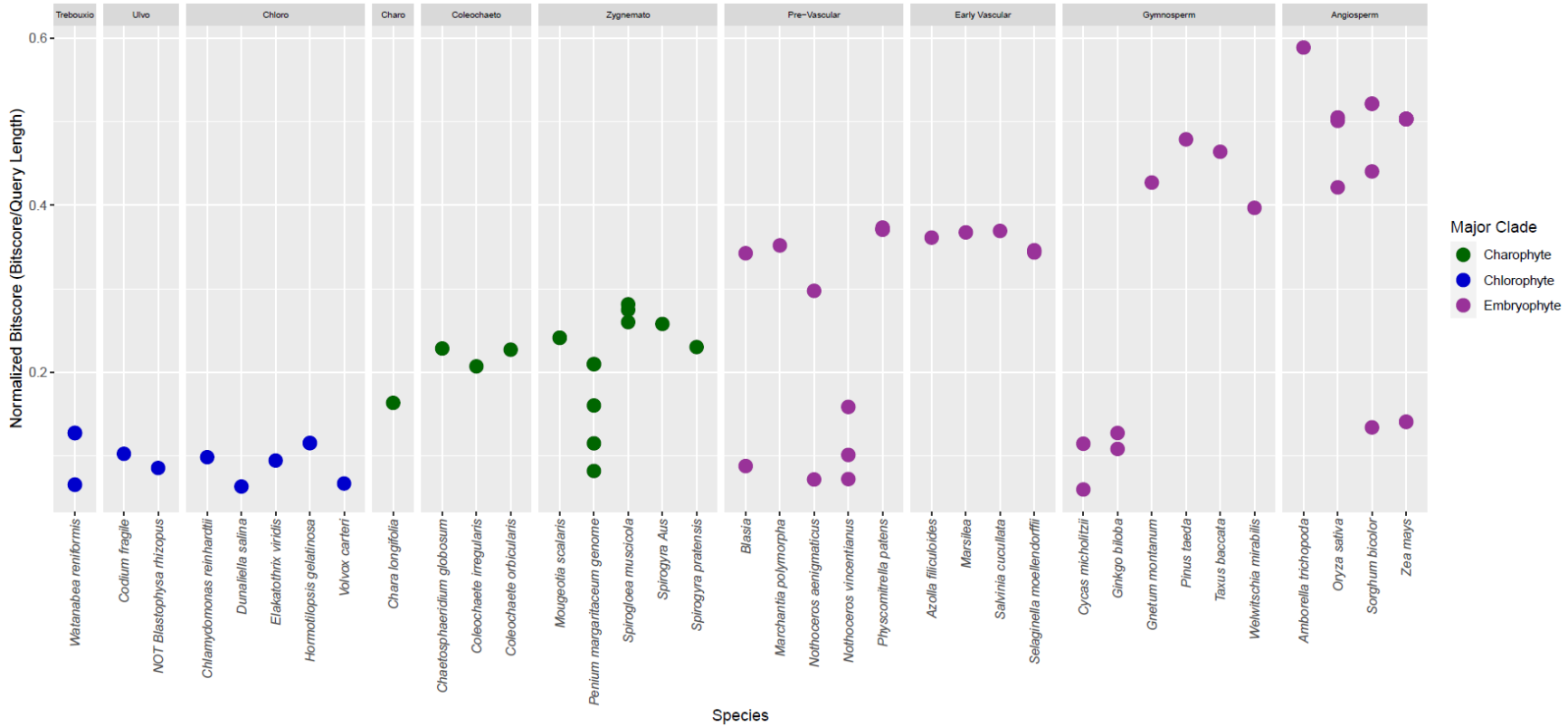


Figure 17: BLAST search of the dataset with the EIN2 signaling domain of the *Arabidopsis thaliana* sequence. Clade abbreviations going from left to right: Trebouxio=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 18: EIN3 Gene Search

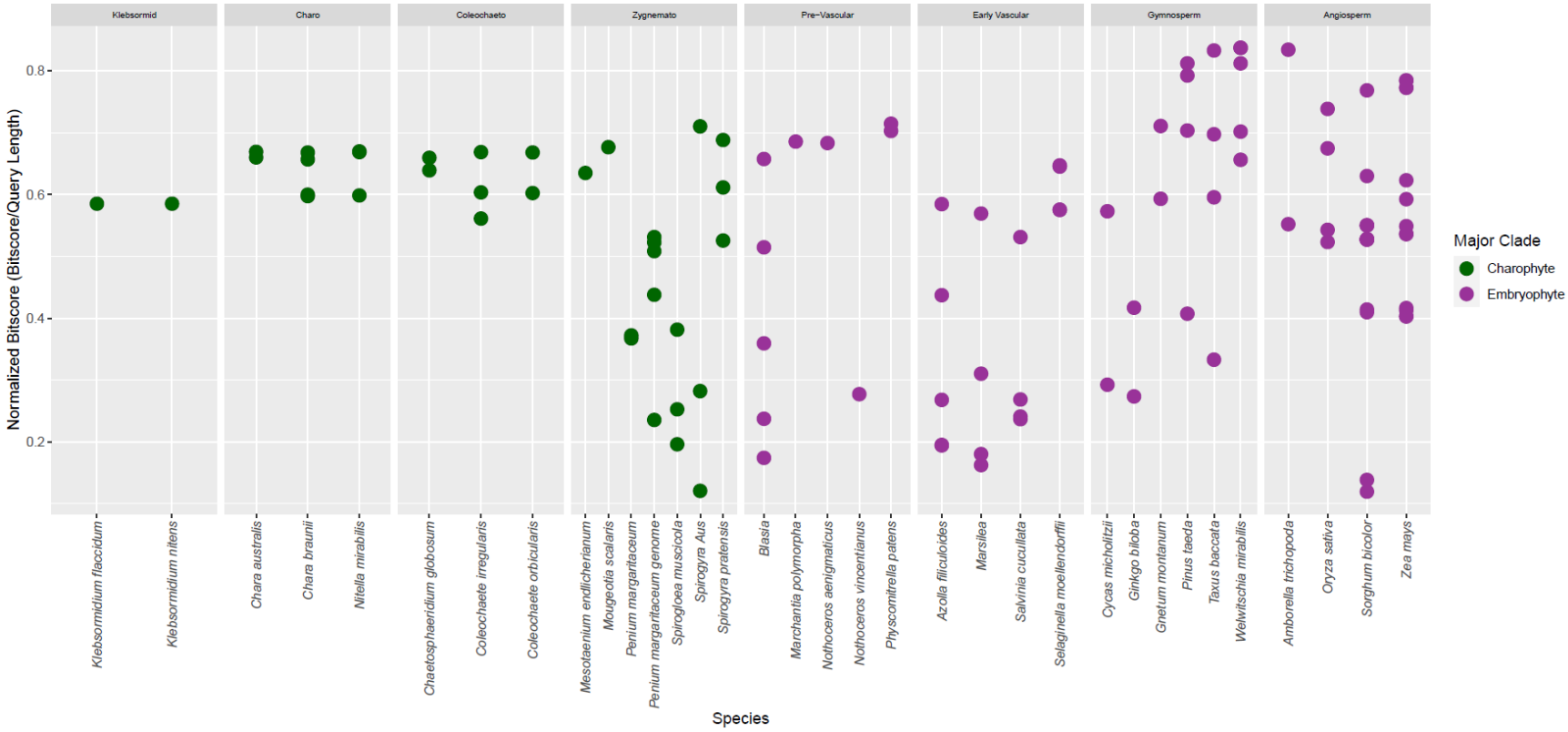


Figure 18: BLAST search of the dataset with the EIN3 sequence with the *Arabidopsis thaliana* sequence. Clade abbreviations going from left to right: Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 19: ETR1 Ethylene Binding Domain Search

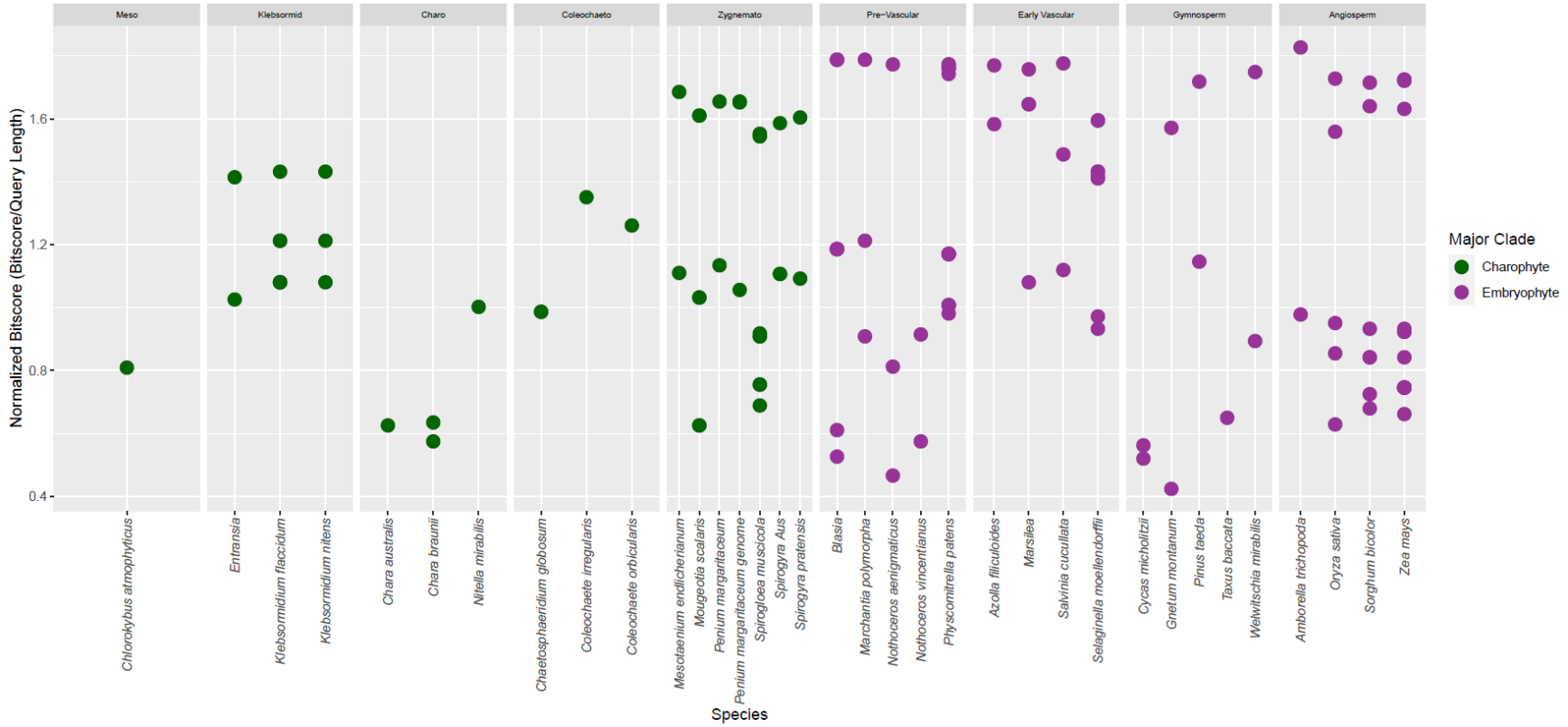


Figure 19: BLAST search of the dataset with the ETR1 ethylene binding domain sequence of the *Arabidopsis thaliana* sequence. Clade abbreviations going from left to right: Meso=Mesostigmatophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 20: NHX Gene Family Searches

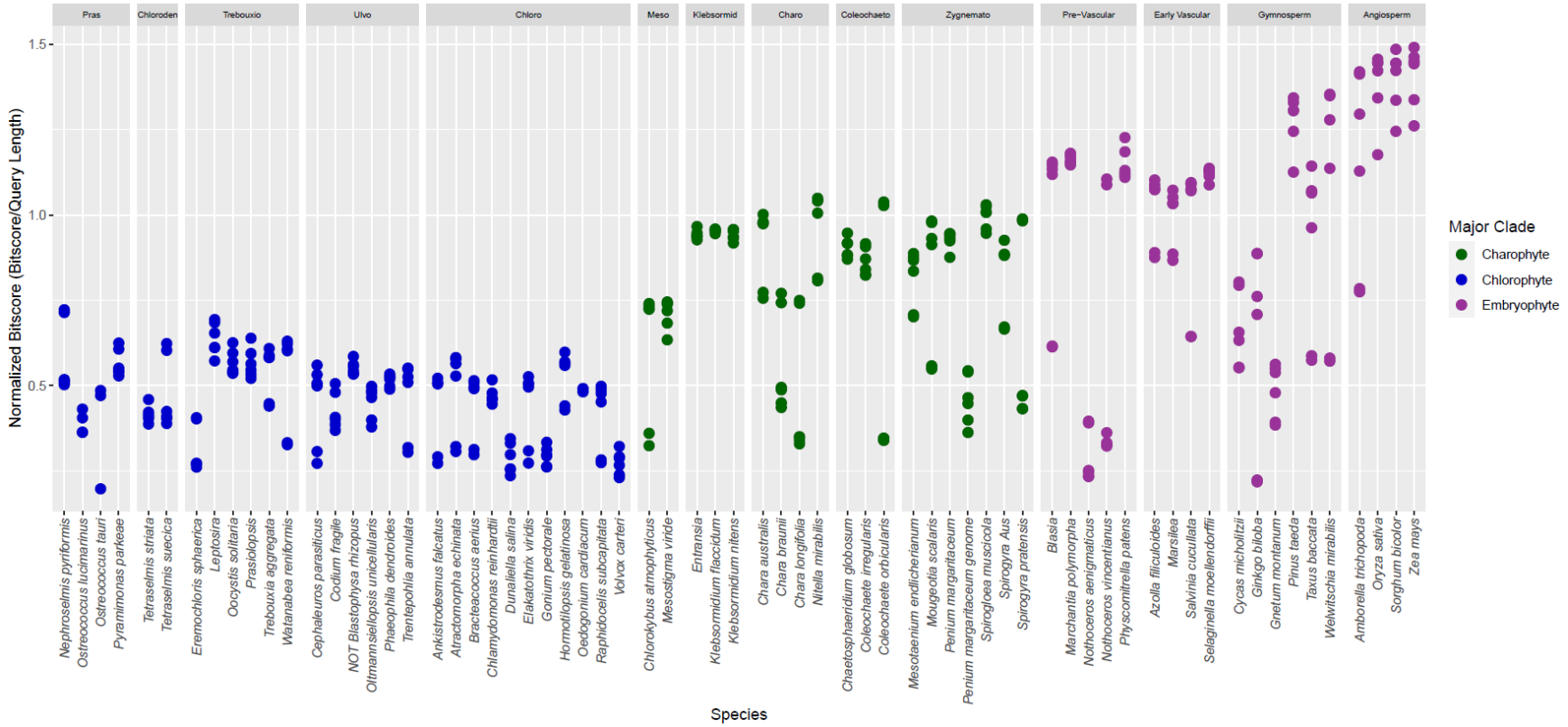


Figure 20: BLAST search of the dataset with the NHX gene family using *Arabidopsis thaliana* sequences (n=6). To reduce clutter, only the best hit for each individual gene in the family is retained for each taxa. Clade abbreviations going from left to right: Pras=Prasinophytes, Chloroden=Chlorodendrophyceae, Trebouxio=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, Meso=Mesostigmatophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 21: CHX Gene Family Searches

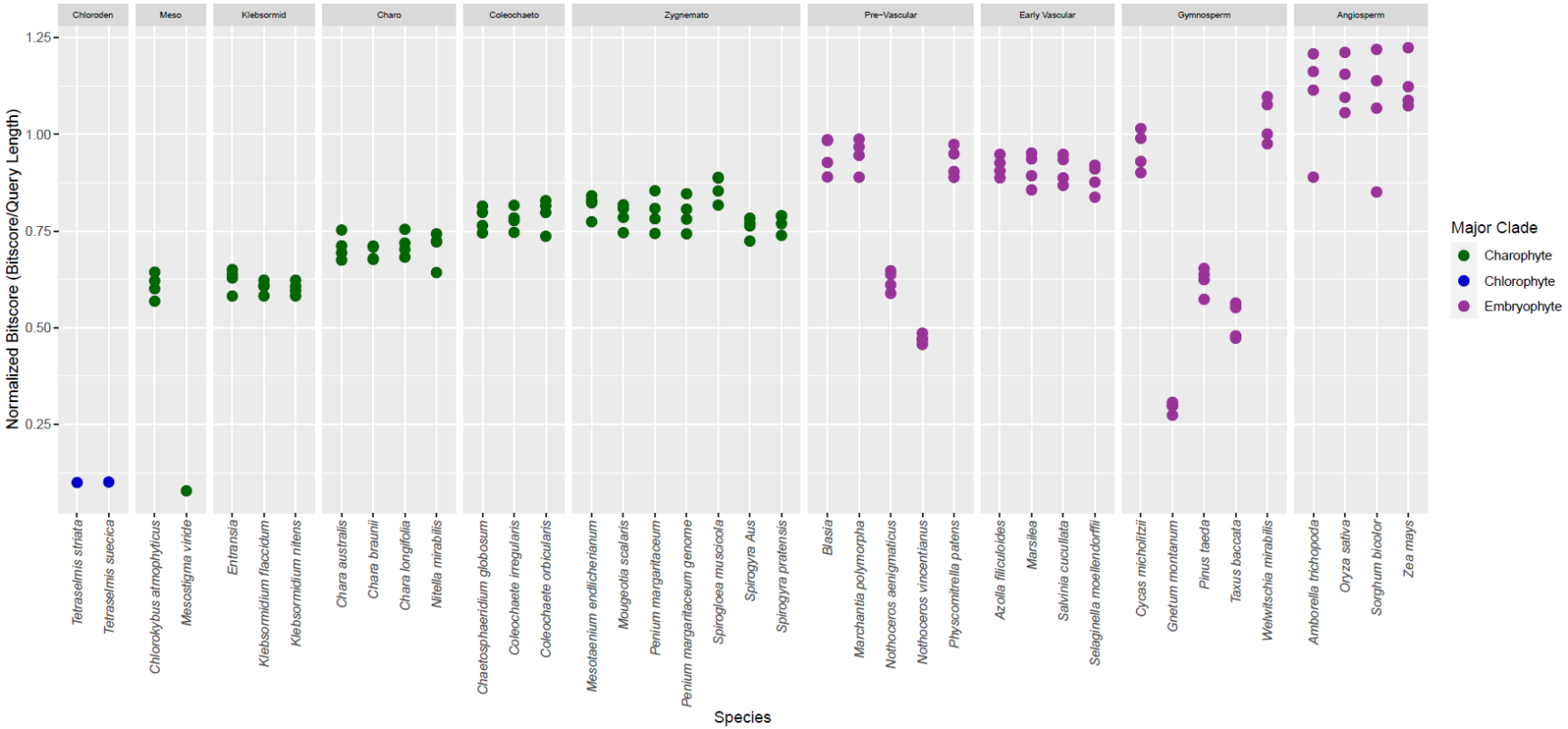


Figure 21: BLAST search of the dataset with the CHX gene family using *Arabidopsis thaliana* sequences (n=4). To reduce clutter, only the best hit for each individual gene in the family is retained for each taxa. Clade abbreviations going from left to right: Chloroden=Chlorodendrophyceae, Meso=Mesostigmatophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 22: KEA Gene Family Searches

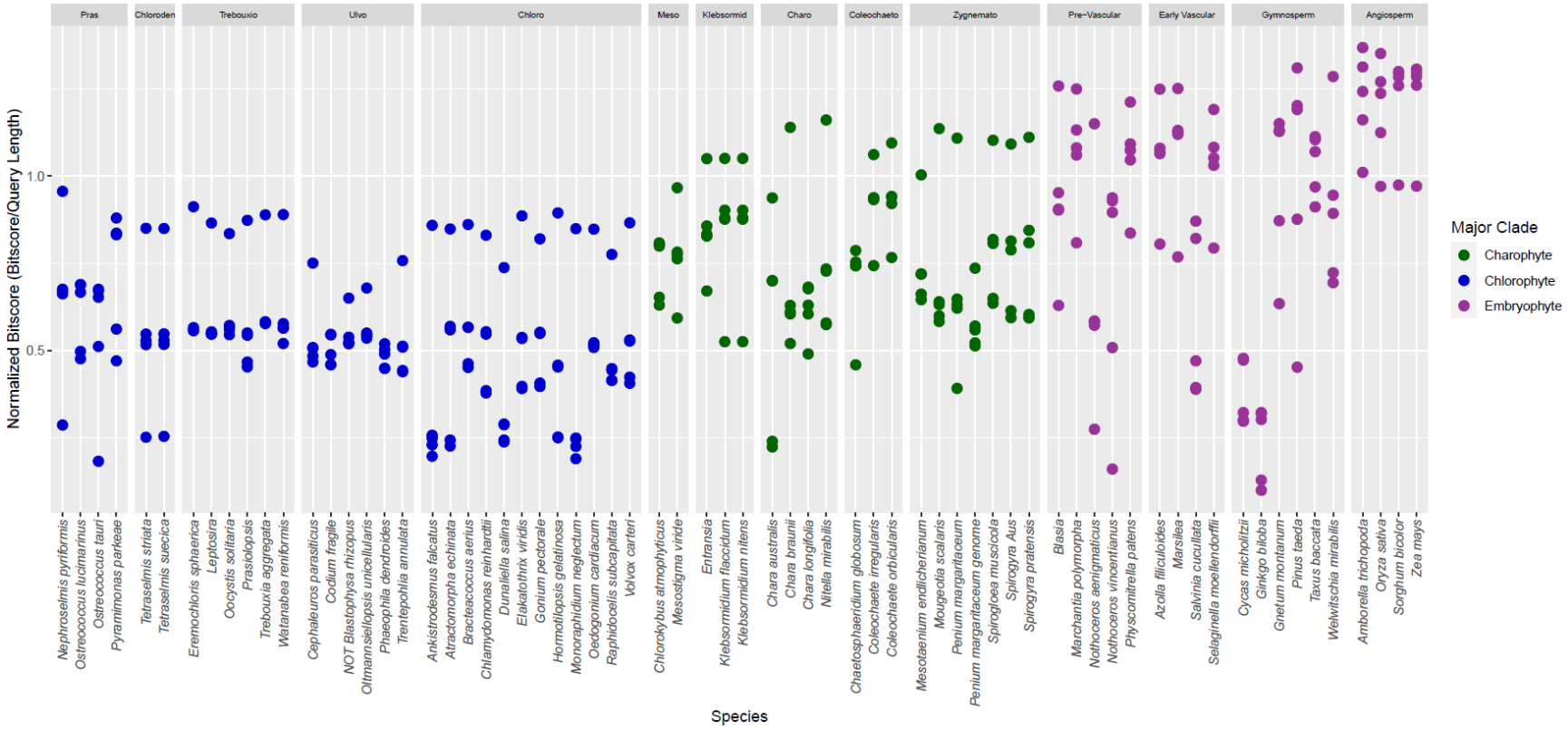


Figure 22: BLAST search of the dataset with the KEA gene family using *Arabidopsis thaliana* sequences (n=6). To reduce clutter, only the best hit for each individual gene in the family is retained for each taxa. Clade abbreviations going from left to right: Pras=Prasinophytes, Chloroden=Chlorodendrophyceae, Trebouxio=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, Meso=Mesostigmatophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Figure 23: AAP Gene Family Searches

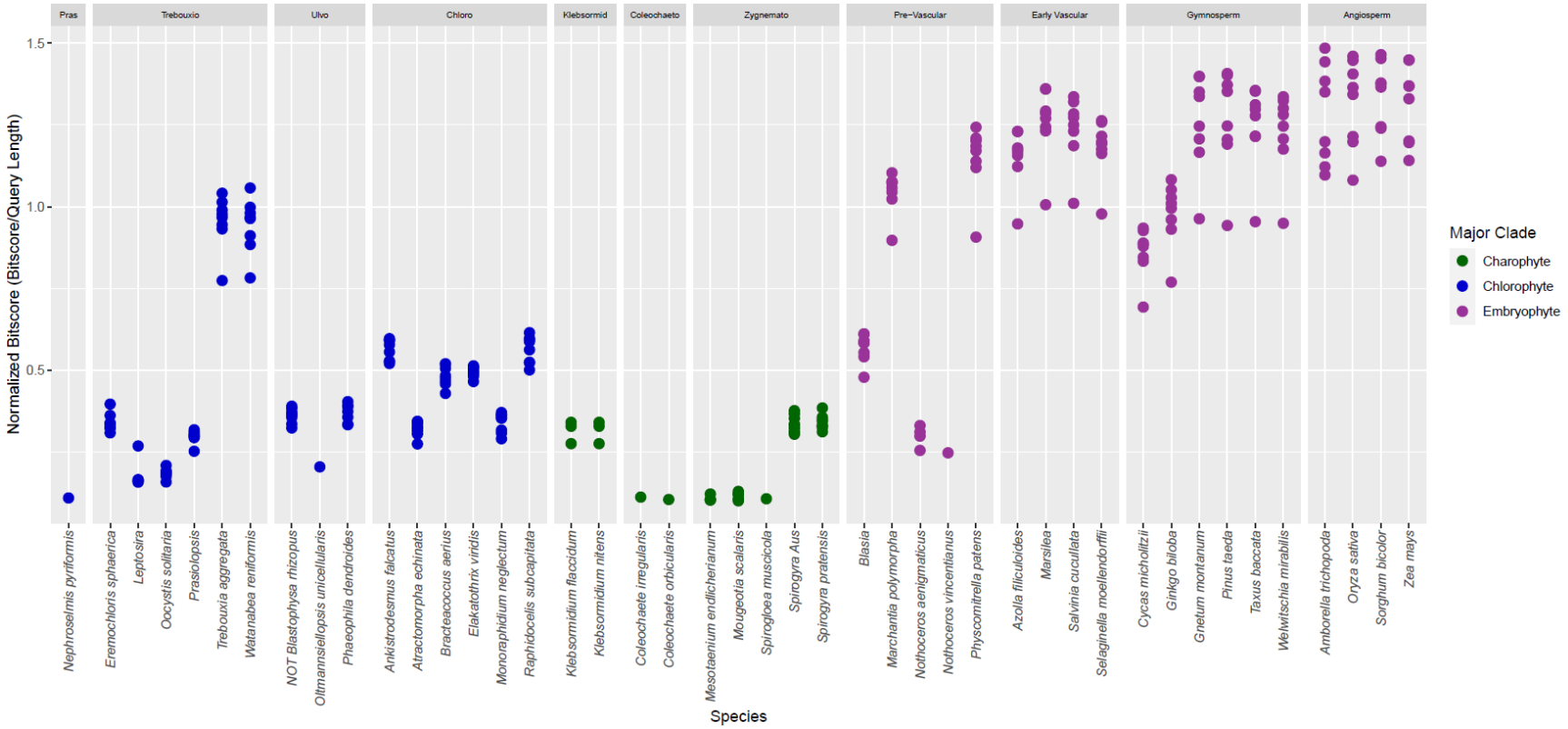
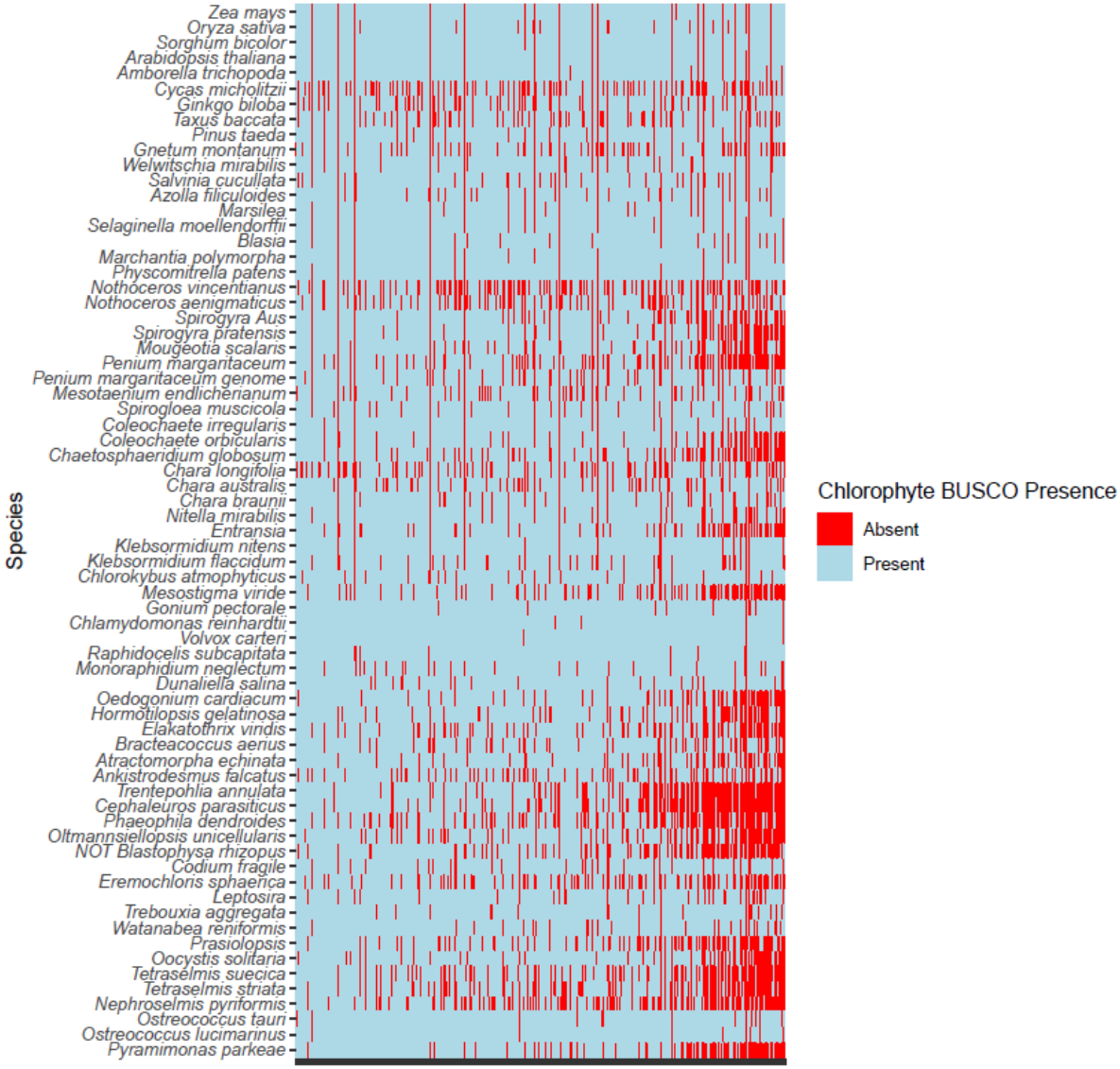
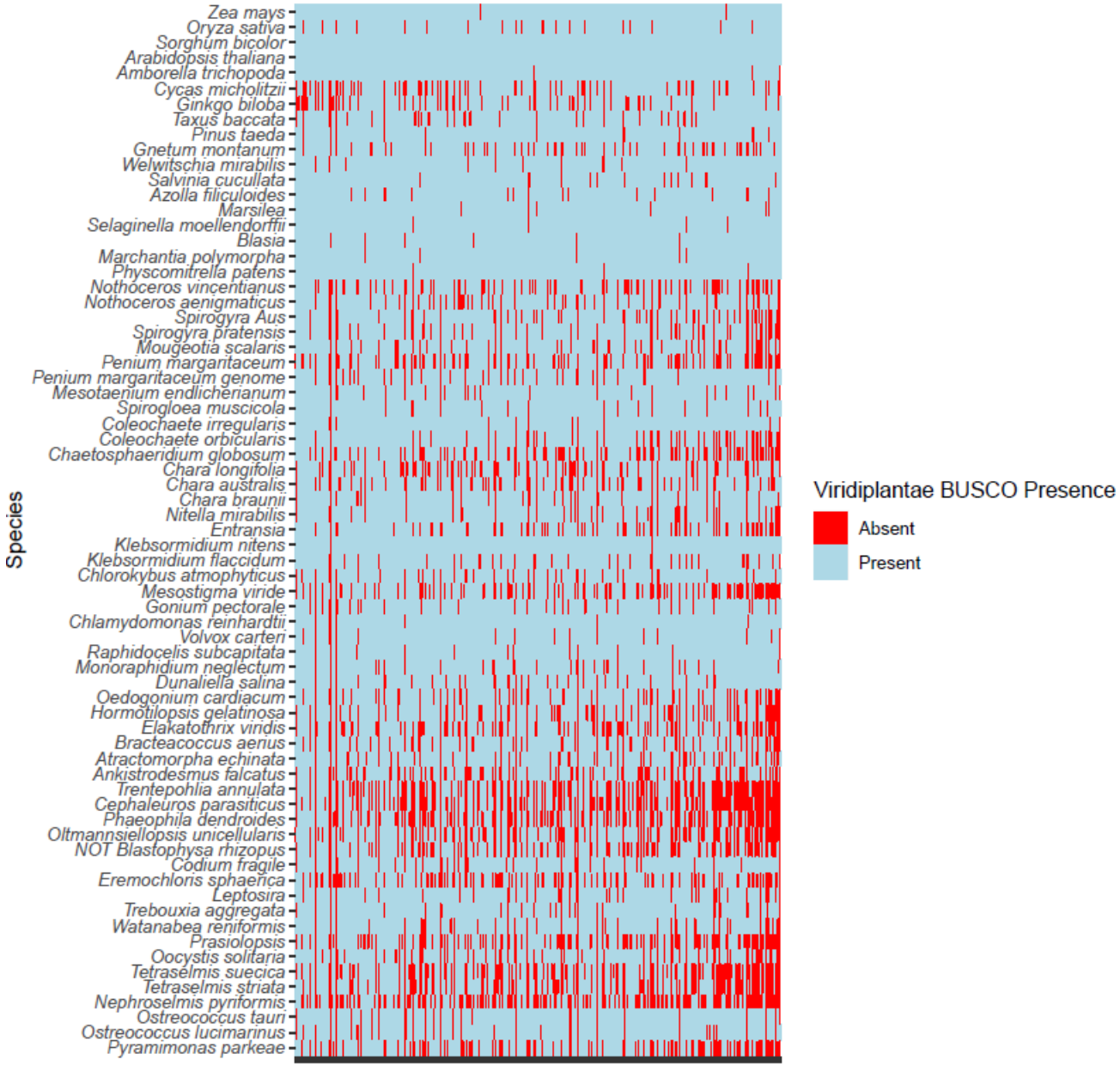
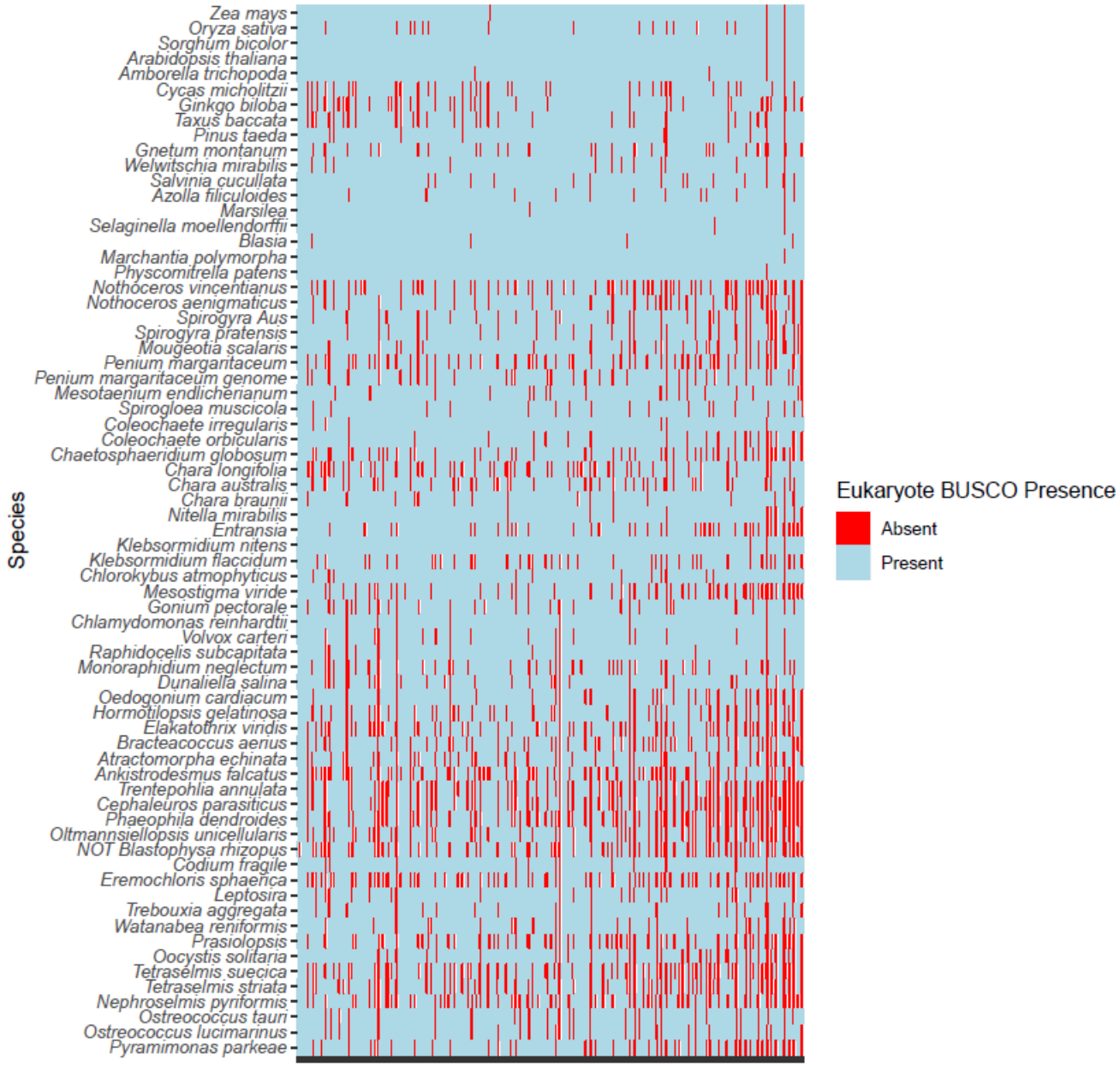


Figure 23: BLAST search of the dataset with the AAP gene family using *Arabidopsis thaliana* sequences (n=8). To reduce clutter, only the best hit for each individual gene in the family is retained for each taxa. Clade abbreviations going from left to right: Pras=Prasinophytes, Trebouxio=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, Klebsormid=Klebsormidiophyceae, Charo=Charophyceae, Coleochaeto=Coleochaetophyceae, Zygnemato=Zygnematophyceae.

Appendix A: Presence/Absence Matrices across the Chlorophyta (1519 genes), Viridiplantate (425 genes), and Eukaryota (255 genes) BUSCO Databases




Species
Zea mays
Oryza sativa
Sorghum bicolor
Arabidopsis thaliana
Amborella trichopoda
Cycas micholitzii
Ginkgo biloba
Taxus baccata
Pinus taeda
Gnetum montanum
Welwitschia mirabilis
Salvinia cucullata
Azolla filiculoides
Marsilea
Selaginella moellendorffii
Blasia
Marchantia polymorpha
Physcomitrella patens
Nothoceros vincentianus
Nothoceros aenigmaticus
Spirogyra Aus
Spirogyra pratensis
Mougeotia scalaris
Penium margaritaceum
Penium margaritaceum genome
Mesotaenium endlicherianum
Spirogloea muscicola
Coleochaete irregularis
Coleochaete orbicularis
Chaetosphaeridium globosum
Chara longifolia
Chara australis
Chara braunii
Nitella mirabilis
Entransia
Klebsormidium nitens
Klebsormidium flaccidum
Chlorokybus atmophyticus
Mesostigma viride
Gonium pectorale
Chlamydomonas reinhardtii
Volvox carteri
Raphidocelis subcapitata
Monoraphidium neglectum
Dunaliella salina
Oedogonium cardiacum
Hormotilopsis gelatinosa
Elakatothrix viridis
Bracteacoccus aerius
Atractomorpha echinata
Ankistrodesmus falcatus
Trentepohlia annulata
Cephaleuros parasiticus
Phaeophila dendroides
Oltmannsiellopsis unicellularis
NOT Blastophysa rhizopus
Codium fragile
Eremochloris sphaerica
Leptosira
Trebouxia aggregata
Watanabea reniformis
Prasiolopsis
Oocystis solitaria
Tetraselmis suecica
Tetraselmis striata
Nephroselmis pyriformis
Ostreococcus tauri
Ostreococcus lucimarinus
Pyramimonas parkeae
Viridiplantae BUSCO Presence
Absent
Present


Species
Zea mays
Oryza sativa
Sorghum bicolor
Arabidopsis thaliana
Amborella trichopoda
Cycas micholitzii
Ginkgo biloba
Taxus baccata
Pinus taeda
Gnetum montanum
Welwitschia mirabilis
Salvinia cucullata
Azolla filiculoides
Marsilea
Selaginella moellendorffii
Blasia
Marchantia polymorpha
Physcomitrella patens
Nothoceros vincentianus
Nothoceros aenigmaticus
Spirogyra Aus
Spirogyra pratensis
Mougeotia scalaris
Penium margaritaceum
Penium margaritaceum genome
Mesotaenium endlicherianum
Spirogloea muscicola
Coleochaete irregularis
Coleochaete orbicularis
Chaetosphaeridium globosum
Chara longifolia
Chara australis
Chara braunii
Nitella mirabilis
Entransia
Klebsormidium nitens
Klebsormidium flaccidum
Chlorokybus atmophyticus
Mesostigma viride
Gonium pectorale
Chlamydomonas reinhardtii
Volvox carteri
Raphidocelis subcapitata
Monoraphidium neglectum
Dunaliella salina
Oedogonium cardiacum
Hormotilopsis gelatinosa
Elakatothrix viridis
Bracteacoccus aerius
Atractomorpha echinata
Ankistrodesmus falcatus
Trentepohlia annulata
Cephaleuros parasiticus
Phaeophila dendroides
Oltmannsiellopsis unicellularis
NOT Blastophysa rhizopus
Codium fragile
Eremochloris sphaerica
Leptosira
Trebouxia aggregata
Watanabea reniformis
Prasiolopsis
Oocystis solitaria
Tetraselmis suecica
Tetraselmis striata
Nephroselmis pyriformis
Ostreococcus tauri
Ostreococcus lucimarinus
Pyramimonas parkeae
Eukaryote BUSCO Presence
Absent
Present

Appendix B: BUSCO ID and function of genes missing from each of the three BUSCO databases for every taxa in each of the three major clades (chlorophyte, charophyte, and embryophyte)

Chlorophytes

BUSCO ID	Database	Function
10857at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones; cellular protein modification process
11445at3041	Chlorophyte	Coenzyme transport and metabolism; Inorganic ion transport and metabolism; Translation, ribosomal structure and biogenesis; Energy production and conversion; nucleotide binding; molybdopterin synthase activity; Mo-molybdopterin cofactor biosynthetic process; cytosol; cytoplasm
11795at3041	Chlorophyte	metal ion binding; chaperone-mediated protein transport; mitochondrion
21163at33090	Viridiplantae	Translation, ribosomal structure and biogenesis; Carbohydrate transport and metabolism; Cell cycle control, cell division, chromosome partitioning; ribonuclease activity; starch binding; RNA phosphodiester bond hydrolysis
22958at33090	Viridiplantae	Posttranslational modification, protein turnover, chaperones; Cell wall/membrane/envelope biogenesis; Signal transduction mechanisms; membrane
215487at33090	Viridiplantae	DNA repair; negative regulation of double-strand break repair via homologous recombination; maintenance of rDNA

Charophytes

BUSCO ID	Database	Function
3069at3041	Chlorophyte	Transcription; Translation, ribosomal structure and biogenesis; Replication, recombination and repair
3558at3041	Chlorophyte	Energy production and conversion; Amino acid transport and metabolism; Posttranslational modification, protein turnover, chaperones
7366at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones
7423at3041	Chlorophyte	Energy production and conversion; metal ion binding; nucleus
8010at3041	Chlorophyte	Secondary metabolites biosynthesis, transport and catabolism; Lipid transport and metabolism; Transcription
8260at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; Transcription; nucleic acid binding
8373at3041	Chlorophyte	Signal transduction mechanisms; Intracellular trafficking, secretion, and vesicular transport; Cell cycle control, cell division, chromosome partitioning; response to Karrikin; granum assembly; chloroplast; chloroplast stroma; chloroplast envelope
8778at3041	Chlorophyte	Translation, ribosomal structure and biogenesis
8863at3041	Chlorophyte	Translation, ribosomal structure and biogenesis
8871at3041	Chlorophyte	integral component of membrane; membrane
8898at3041	Chlorophyte	Signal transduction mechanisms; Transcription; Cell motility
8920at3041	Chlorophyte	Signal transduction mechanisms; Secondary metabolites biosynthesis, transport and catabolism; integral component of membrane; membrane
8925at3041	Chlorophyte	Transcription; Cell cycle control, cell division, chromosome partitioning; Defense mechanisms; sequence-specific DNA binding
8977at3041	Chlorophyte	Replication, recombination and repair; GINS complex
9109at3041	Chlorophyte	Energy production and conversion; Lipid transport and metabolism; Amino acid transport and metabolism; Coenzyme transport and metabolism; glycine decarboxylation via glycine cleavage system; mitochondrion
9586at3041	Chlorophyte	N/A
9588at3041	Chlorophyte	Mitochondrion; mitochondrial inner membrane
9759at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; Posttranslational modification, protein turnover, chaperones; Amino acid transport and metabolism; H2A histone acetyltransferase activity; H4 histone acetyltransferase activity; histone H2A acetylation; histone H4 acetylation

9788at3041	Chlorophyte	Translation, ribosomal structure and biogenesis
10197at3041	Chlorophyte	N/A
10934at3041	Chlorophyte	Translation, ribosomal structure and biogenesis
11022at3041	Chlorophyte	Transcription; Signal transduction mechanisms
11481at3041	Chlorophyte	sister chromatid cohesion
11894at3041	Chlorophyte	N/A
12521at3041	Chlorophyte	Membrane; integral component of membrane
21163at33090	Viridiplantae	Translation, ribosomal structure and biogenesis; Carbohydrate transport and metabolism; Cell cycle control, cell division, chromosome partitioning; ribonuclease activity; starch binding; RNA phosphodiester bond hydrolysis

Embryophytes

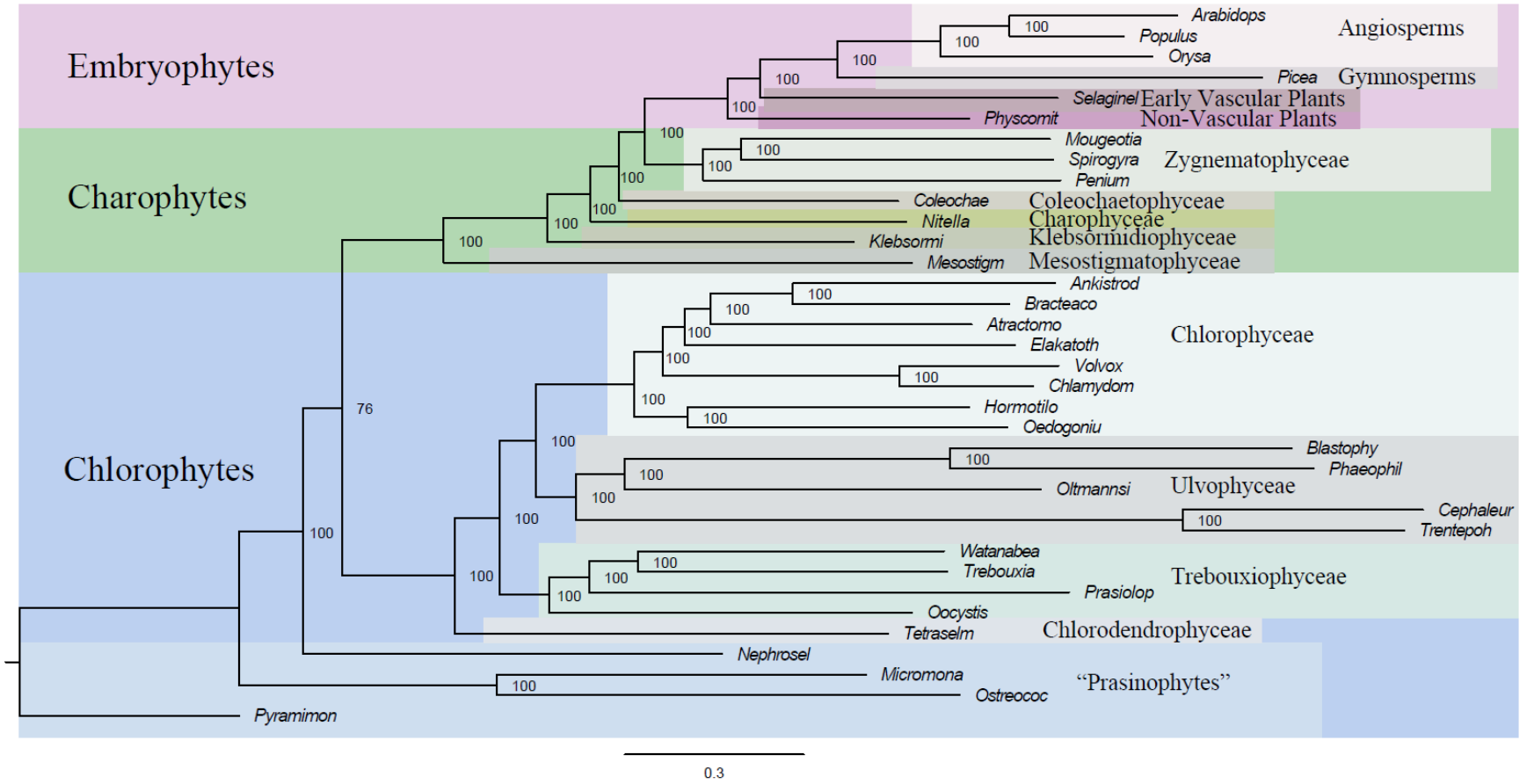
BUSCO ID	Database	Function
91at3041	Chlorophyte	Energy production and conversion; Signal transduction mechanisms; Nucleotide transport and metabolism; catalytic activity; flavin adenine dinucleotide binding; iron-sulfur cluster binding; FAD binding; oxidoreductase activity; oxidation-reduction process
652at3041	Chlorophyte	Replication, recombination and repair; Transcription; Defense mechanisms; Translation, ribosomal structure and biogenesis; ATP binding; nucleic acid binding; helicase activity; hydrolase activity; nucleotide binding
766at3041	Chlorophyte	Inorganic ion transport and metabolism; posttranslational modification, protein turnover, chaperones; Translation, ribosomal structure and biogenesis; motor activity; ATP binding; microtubule motor activity; nucleotide binding; microtubule binding; microtubule-based movement; microtubule
948at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; Posttranslational modification, protein turnover, chaperones; Mobilome: prophages, transposons; Replication, recombination and repair; RNA ligase activity; ligase activity; ATP binding; nucleotide binding; RNA ligase (ATP) activity; metal ion binding; tRNA processing; tRNA splicing, via endonucleolytic cleavage and ligation; tRNA-splicing ligase complex
1271at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; Signal transduction mechanisms; Inorganic ion transport and metabolism; GTPase activity; GTP binding
1997at3041	Chlorophyte	N/A
2520at3041	Chlorophyte	Defense mechanisms; Translation, ribosomal structure and biogenesis
2855at3041	Chlorophyte	N/A
3375at3041	Chlorophyte	Intracellular trafficking, secretion, and vesicular transport; Translation, ribosomal structure and biogenesis; asparaginase activity
3558at3041	Chlorophyte	Energy production and conversion; Amino acid transport and metabolism; Posttranslational modification, protein turnover, chaperones
3562at3041	Chlorophyte	Replication, recombination and repair; Energy production and conversion; Lipid transport and metabolism; Translation, ribosomal structure and biogenesis; alpha-1,6-mannosyltransferase activity; mannosylation
3587at3041	Chlorophyte	Signal transduction mechanisms; Transcription; Amino acid transport and metabolism; catalytic activity; metal ion binding; cation binding; phosphoprotein phosphatase activity; hydrolase activity

3692at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; Coenzyme transport and metabolism; Cell cycle control, cell division, chromosome partitioning; methylation
3946at3041	Chlorophyte	Signal transduction mechanisms; Coenzyme transport and metabolism; Replication, recombination and repair; ATP binding; nucleotide binding; protein phosphorylation
3998at3041	Chlorophyte	Signal transduction mechanisms; membrane; integral component of membrane
4147at3041	Chlorophyte	protein ubiquitination
4553at3041	Chlorophyte	Signal transduction mechanisms; Cell wall/membrane/envelope biogenesis; Coenzyme transport and metabolism; chlorophyll binding; protein-chromophore linkage; photosynthesis; photosynthesis, light harvesting; chloroplast; thylakoid; photosystem I; photosystem II; plastid; integral component of membrane; membrane; chloroplast thylakoid membrane
4567at3041	Chlorophyte	Defense mechanisms; Amino acid transport and metabolism; Inorganic ion transport and metabolism; nucleotide binding; ATP binding
4754at3041	Chlorophyte	Coenzyme transport and metabolism; chlorophyll binding; protein-chromophore linkage; photosynthesis; photosynthesis, light harvesting; thylakoid; chloroplast; photosystem II; plastid; chloroplast thylakoid membrane; membrane; integral component of membrane; photosystem I
4958at3041	Chlorophyte	Lipid transport and metabolism; Coenzyme transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism; Translation, ribosomal structure and biogenesis; catalytic activity; isomerase activity
5075at3041	Chlorophyte	Coenzyme transport and metabolism; Replication, recombination and repair; Lipid transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism; riboflavin synthase activity
5327at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones; Extracellular structures; Signal transduction mechanisms; protein import into chloroplast stroma; chloroplast inner membrane; chloroplast envelope
5387at3041	Chlorophyte	Translation, ribosomal structure and biogenesis
5444at3041	Chlorophyte	Intracellular trafficking, secretion, and vesicular transport; Replication, recombination and repair; Mobilome: prophages, transposons; ATP binding
5533at3041	Chlorophyte	N/A
5599at3041	Chlorophyte	Signal transduction mechanisms; Carbohydrate transport and metabolism; Intracellular trafficking, secretion, and vesicular transport

5946at3041	Chlorophyte	Lipid transport and metabolism; Coenzyme transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism; hydrolase activity, acting on ester bonds
5994at3041	Chlorophyte	Coenzyme transport and metabolism; Secondary metabolites biosynthesis, transport and catabolism; Signal transduction mechanisms; catalytic activity
6192at3041	Chlorophyte	nucleus
6602at3041	Chlorophyte	Amino acid transport and metabolism; Signal transduction mechanisms; Transcription; nucleotidyltransferase activity; transferase activity
6682at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones; Transcription; Inorganic ion transport and metabolism; oxidation-reduction process; glycerol ether metabolic process; cell
6972at3041	Chlorophyte	N/A
7143at3041	Chlorophyte	Coenzyme transport and metabolism; Replication, recombination and repair; Signal transduction mechanisms; Translation, ribosomal structure and biogenesis; methylation
7366at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones
7588at3041	Chlorophyte	Inorganic ion transport and metabolism; Carbohydrate transport and metabolism; membrane; integral component of membrane
7631at3041	Chlorophyte	Energy production and conversion; Coenzyme transport and metabolism; Nucleotide transport and metabolism; iron-sulfur cluster binding; 2 iron, 2 sulfur cluster binding; electron transport chain
7648at3041	Chlorophyte	Replication, recombination and repair; RNA processing and modification; Signal transduction mechanisms; nucleic acid binding
7903at3041	Chlorophyte	N/A
7963at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones; Transcription
8100at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones; Replication, recombination and repair; metal ion binding
8134at3041	Chlorophyte	Inorganic ion transport and metabolism; Signal transduction mechanisms; Energy production and conversion; solute:proton antiporter activity; transmembrane transport; membrane; integral component of membrane
8989at3041	Chlorophyte	integral component of membrane; membrane
9277at3041	Chlorophyte	N/A
9319at3041	Chlorophyte	N/A
9426at3041	Chlorophyte	Signal transduction mechanisms; Coenzyme transport and metabolism; Carbohydrate transport and metabolism; sulfotransferase activity; integral component of membrane

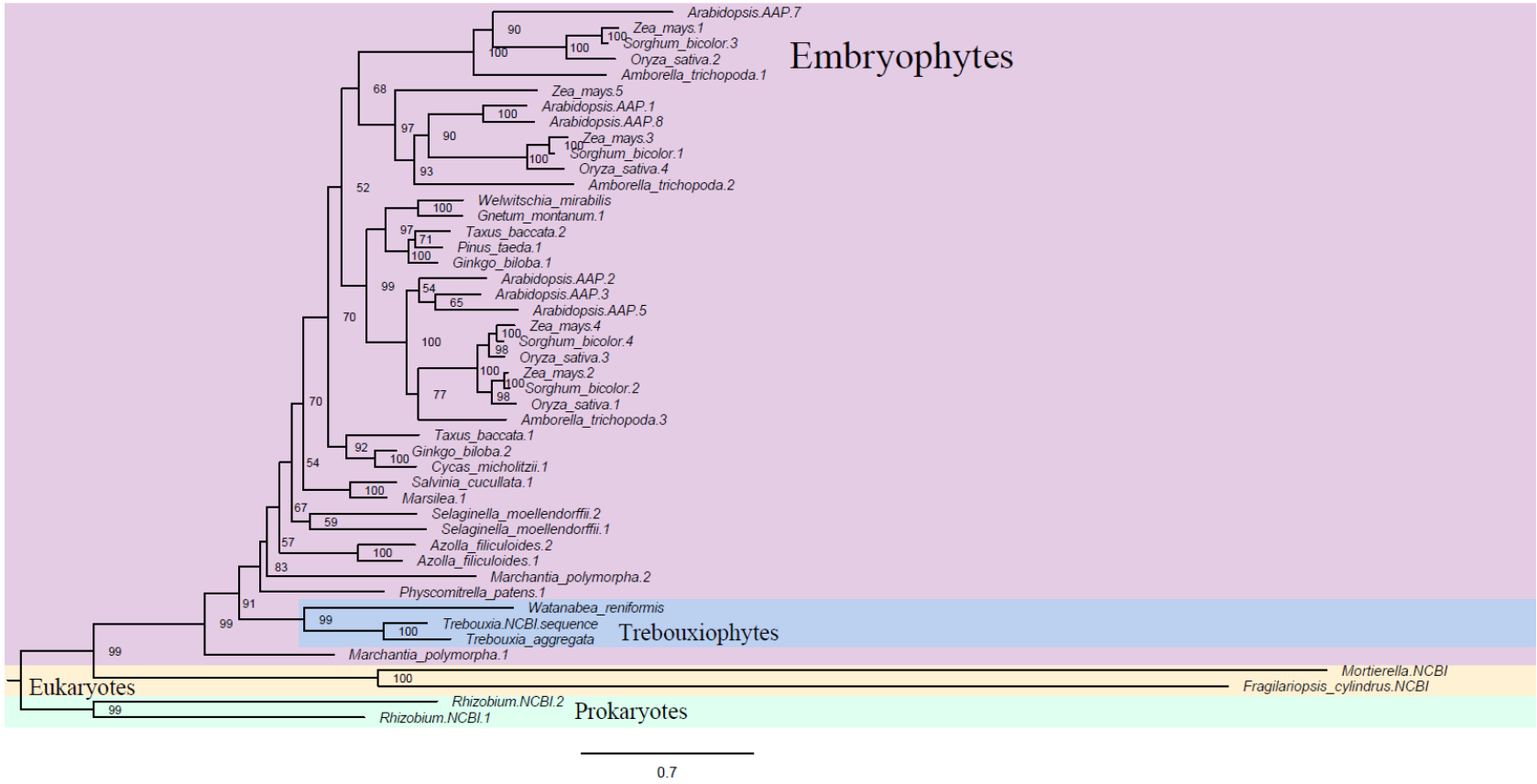
9586at3041	Chlorophyte	N/A
9860at3041	Chlorophyte	Signal transduction mechanisms; Replication, recombination and repair; histone methylation; ESC/E(Z) complex
10175at3041	Chlorophyte	Membrane; integral component of membrane
10182at3041	Chlorophyte	photosystem II assembly
10319at3041	Chlorophyte	Energy production and conversion; Posttranslational modification, protein turnover, chaperones; Amino acid transport and metabolism; Transcription; oxidoreductase activity, acting on the CH-CH group of donors; membrane
10409at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; tRNA splicing, via endonucleolytic cleavage and ligation; tRNA-splicing ligase complex
10517at3041	Chlorophyte	Secondary metabolites biosynthesis, transport and catabolism; Lipid transport and metabolism; Energy production and conversion
10539at3041	Chlorophyte	oxidation-reduction process; membrane; integral component of membrane
11739at3041	Chlorophyte	Translation, ribosomal structure and biogenesis; Energy production and conversion
12455at3041	Chlorophyte	N/A
12560at3041	Chlorophyte	Posttranslational modification, protein turnover, chaperones; Transcription

Appendix C: Phylogenetic Tree from Endymion Cooper's Alignment



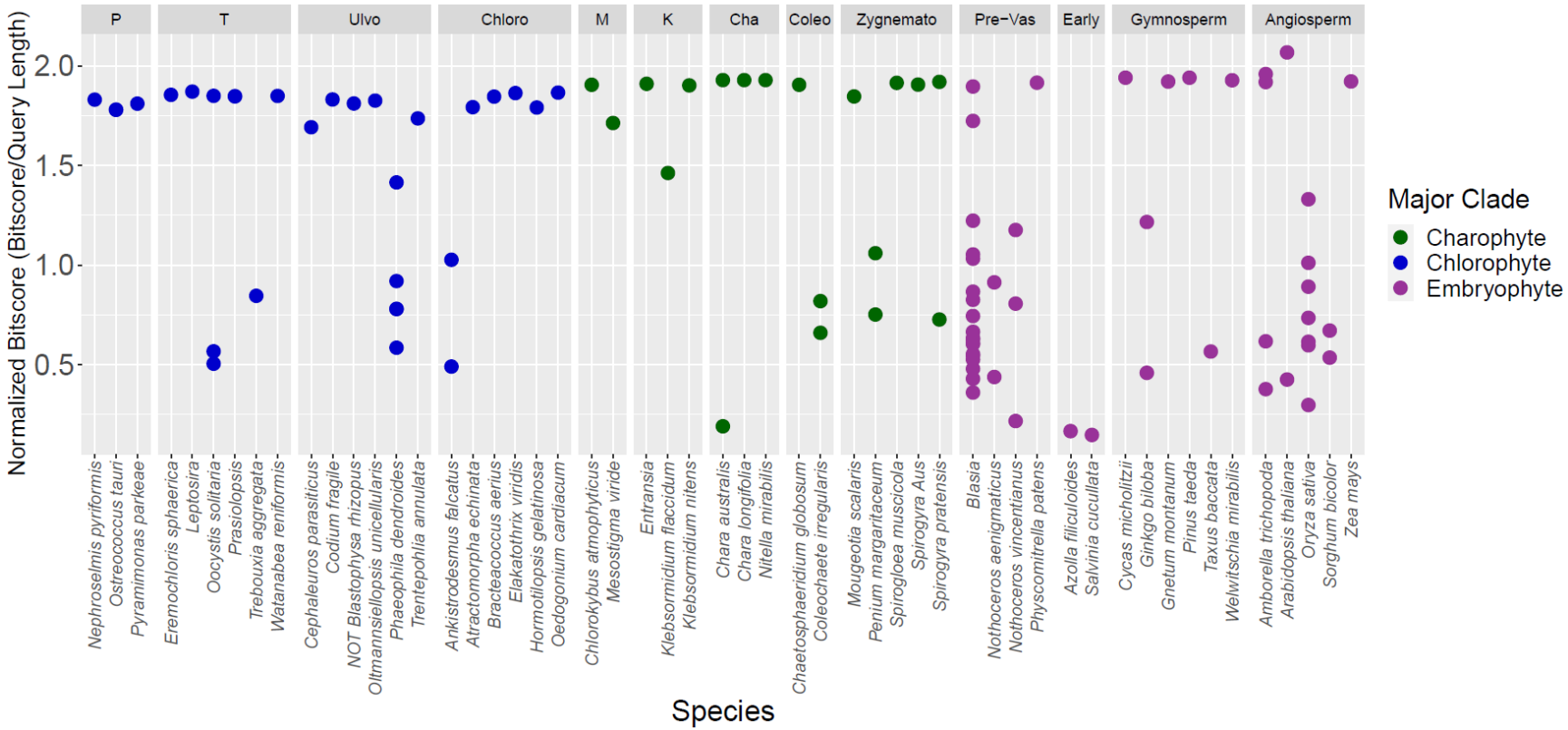
Phylogenetic tree constructed with Endymion Cooper's multiple sequence alignment consisting of 998096 sites from ~1700 orthologous genes and the LG+F+R10 model in IQ-TREE using *Pyramimonas parkeae* as the root and ten thousand ultrafast bootstrap replicates. Bootstrap support is shown at each node in the tree.

Appendix D: Amino Acid Permease Tree



Phylogenetic tree constructed from BLAST hits from the amino acid permease search with information density > 0.75 as well as sequences from another Trebouxiophyte, other eukaryotes, and bacteria from BLAST searches on NCBI. The tree was rooted using sequences from *Rhizobium*, a soil bacteria. Eukaryote sequences are from *Fragilariopsis cylindrus,* a diatom, and *Mortierella,* a fungus. The analysis used the LG+F+G4 model (LG substitution matrix, empirical amino acid frequencies, and four rate heterogeneity parameters drawn from a gamma distribution). Bootstrap support values from ten thousand ultrafast bootstrap replicates are shown at each node.

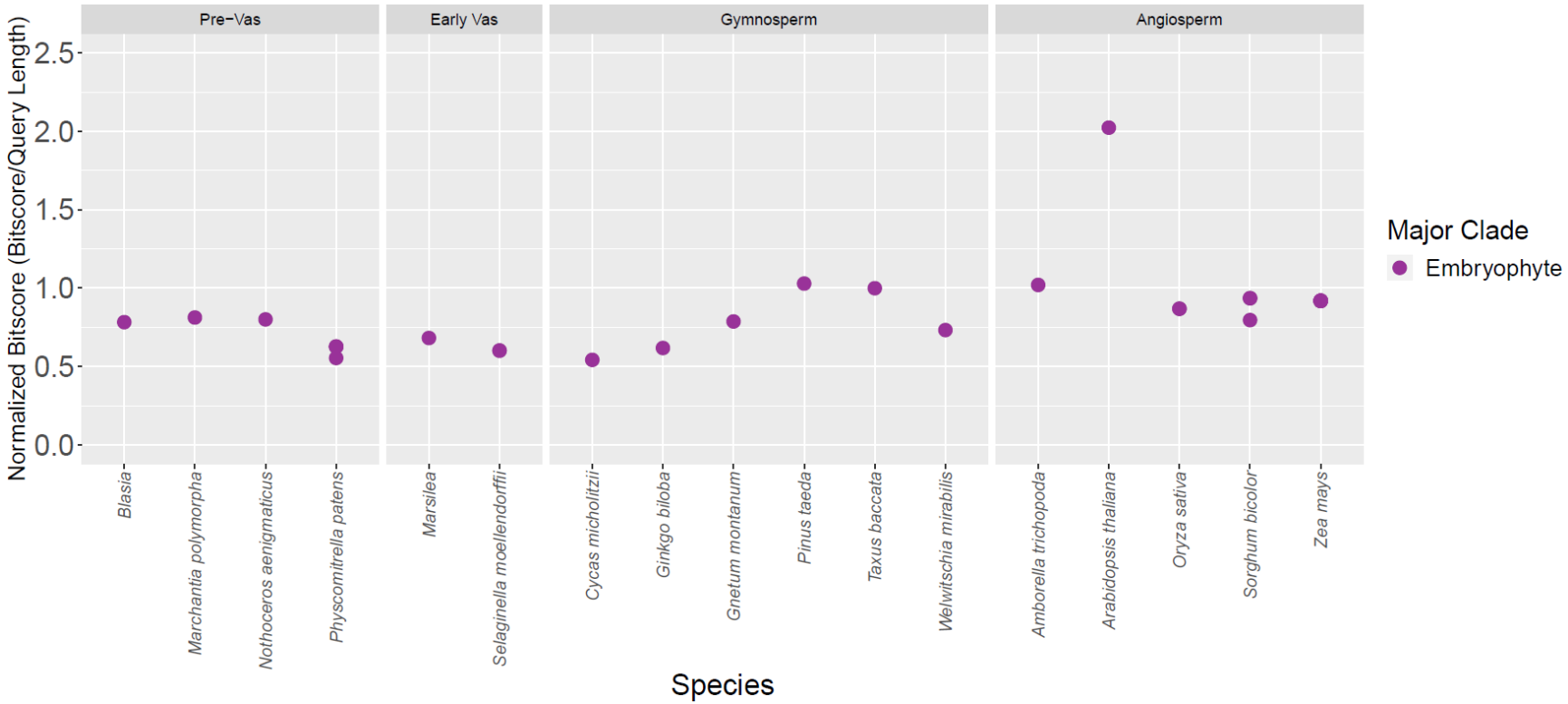
Appendix E: Rubisco BLAST Search



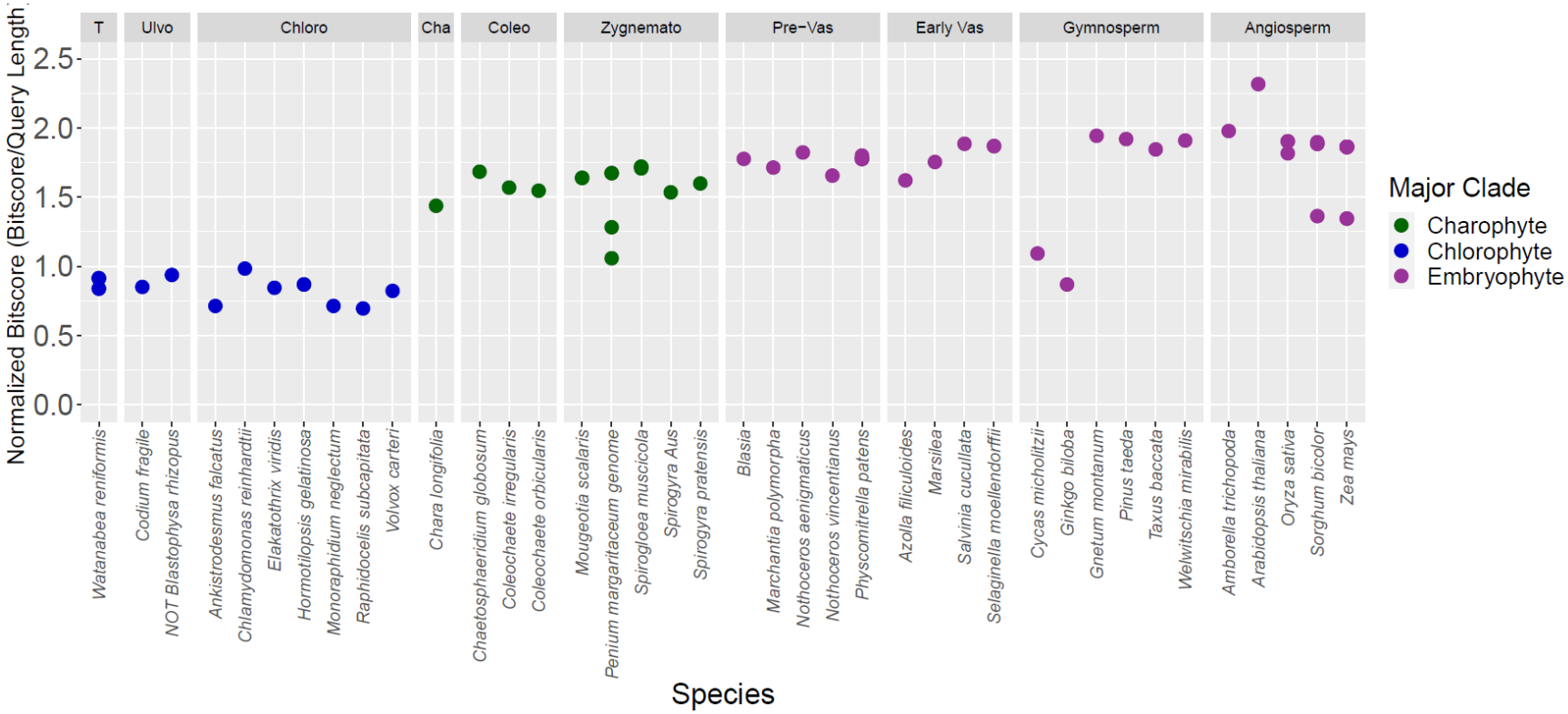
BLAST search of the dataset using the *Arabidopsis thaliana* sequence for the RuBisCO gene. Clade abbreviations going from left to right: P=Prasinophytes, T=Trebouxiophyceae, Ulvo=Ulvophyceae, Chloro=Chlorophyceae, M=Mesostigmatophyceae, K=Klebsormidiophyceae, Cha=Charophyceae, Coleo=Coleochaetophyceae, Zygnemato=Zygnematophyceae, Pre-Vas=Pre-Vascular, Early=Early Vascular

Appendix F: EIN 2 Subdomains

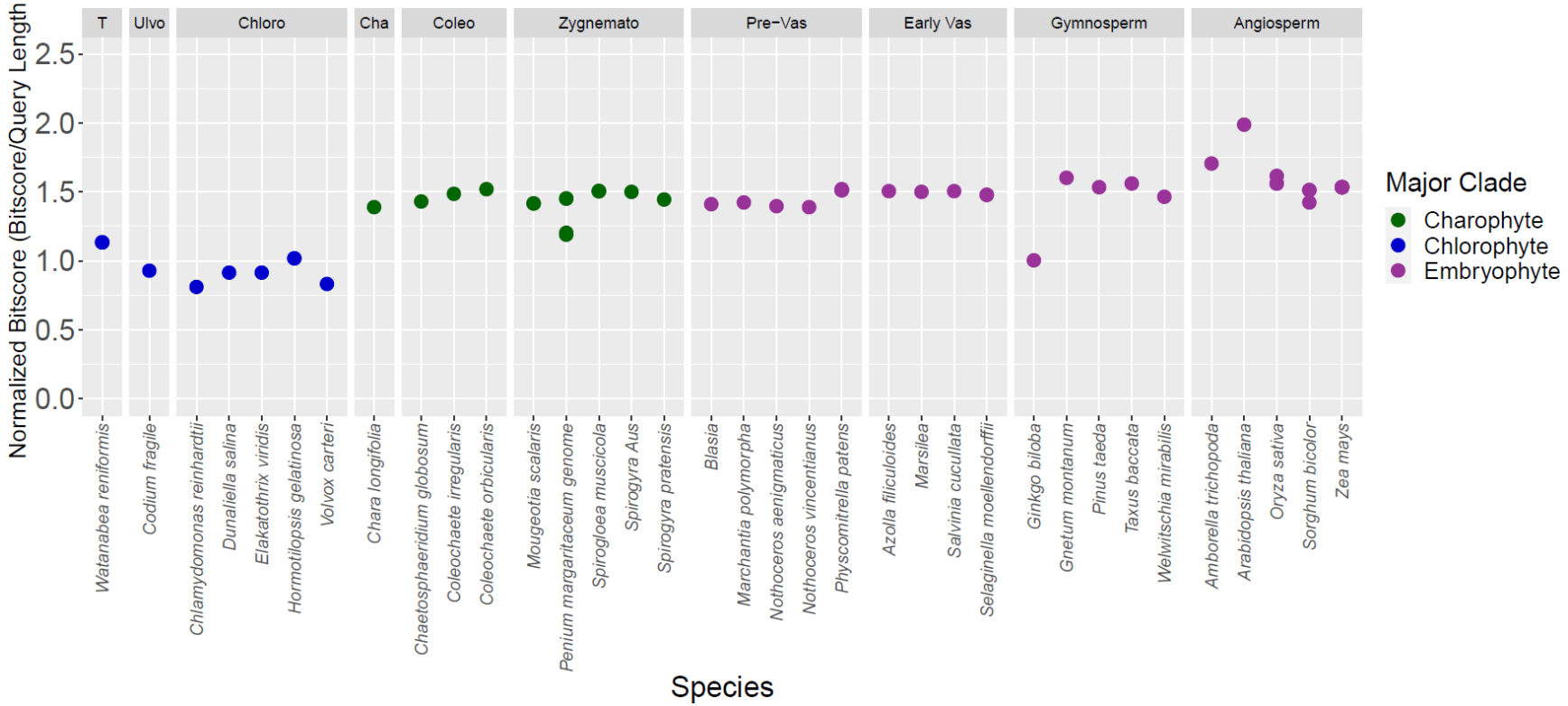
Arabidopsis thaliana Subdomain 1



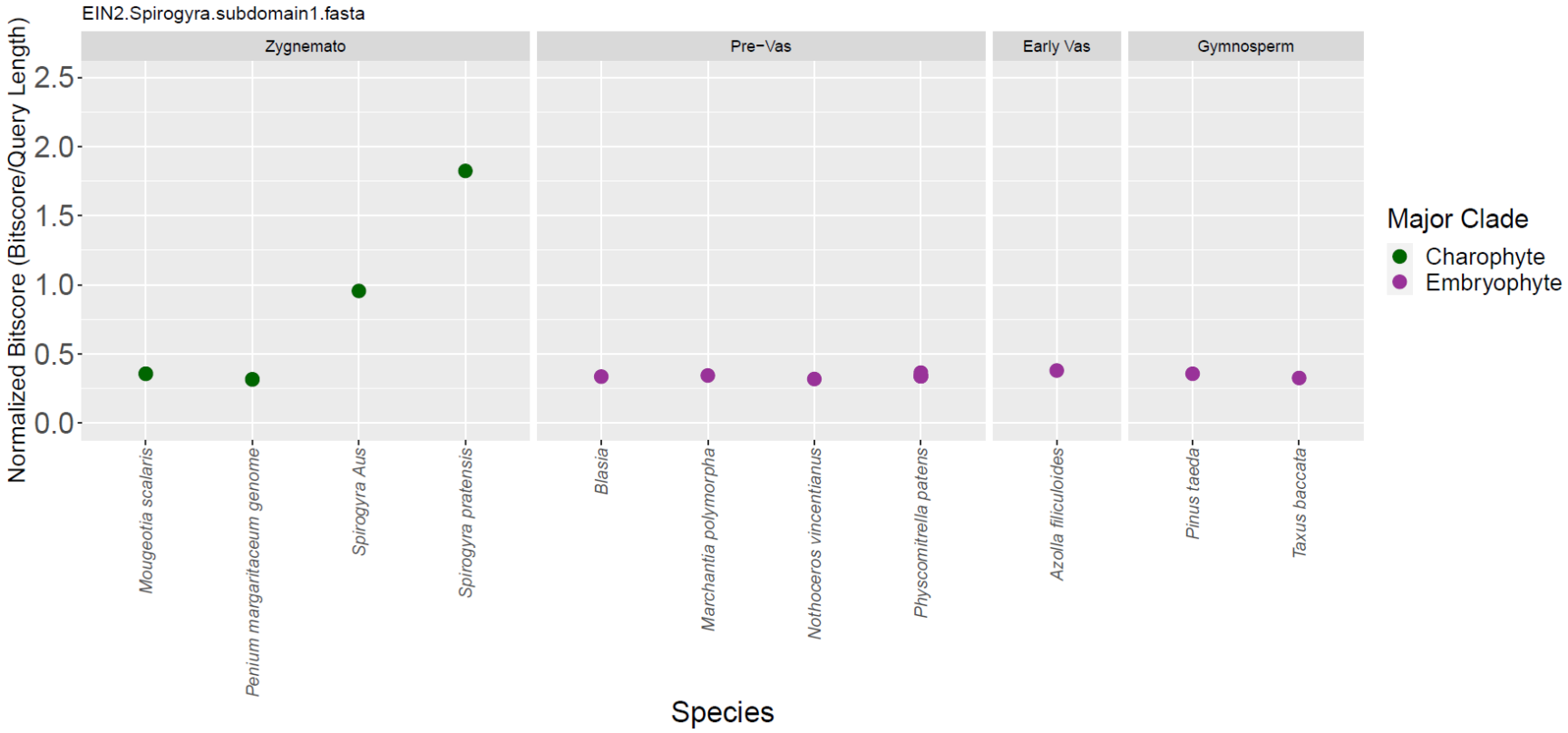
***Arabidopsis thaliana* Subdomain 2**



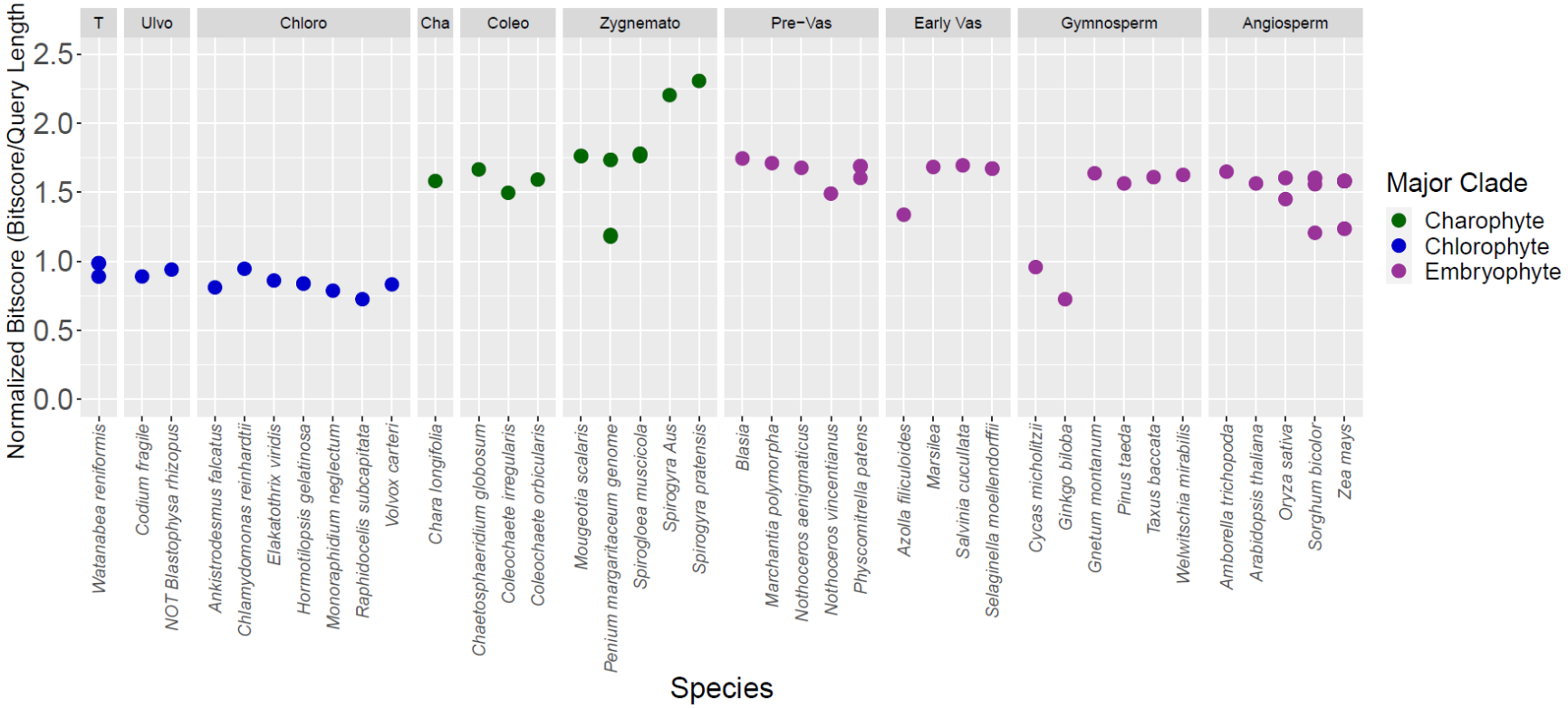
***Arabidopsis thaliana* Subdomain 3**



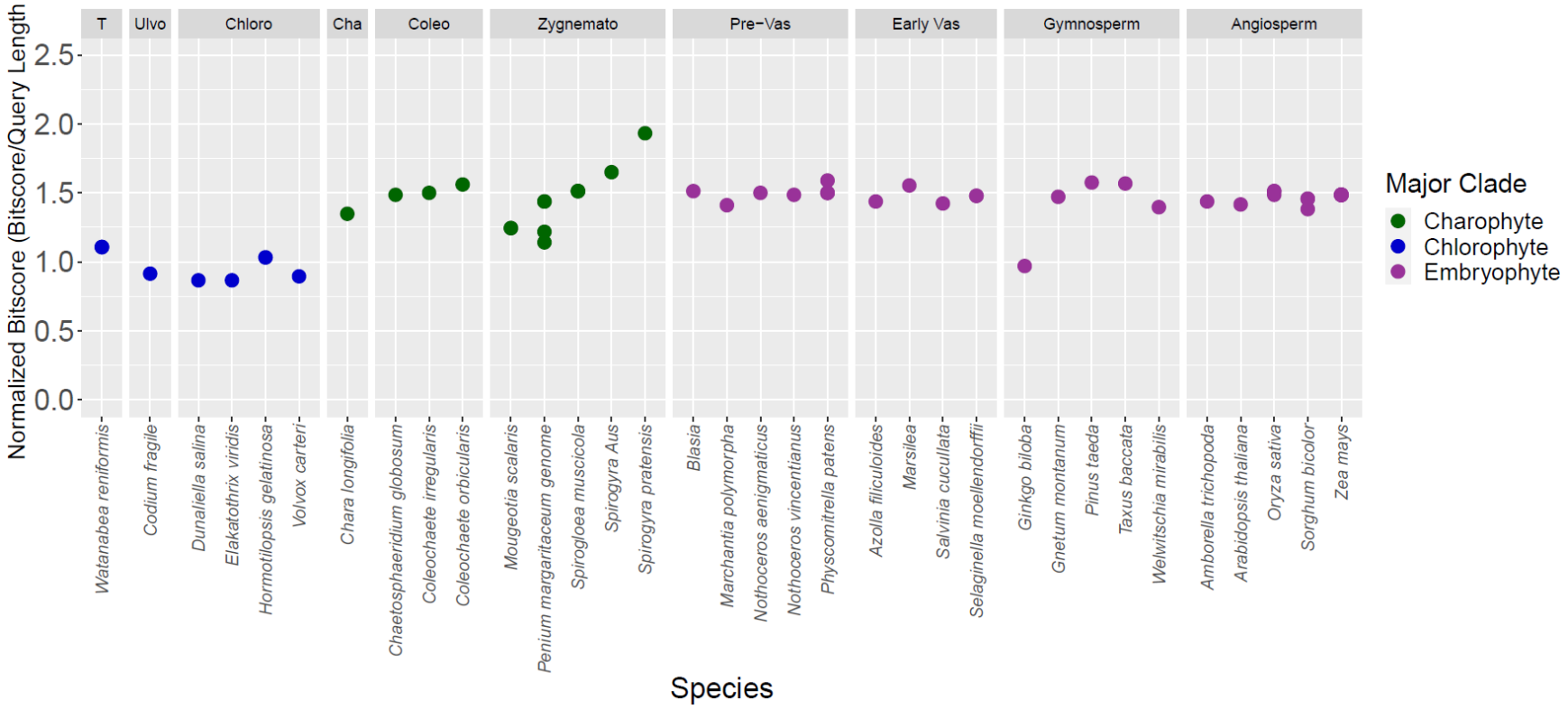
Spirogyra Subdomain 1



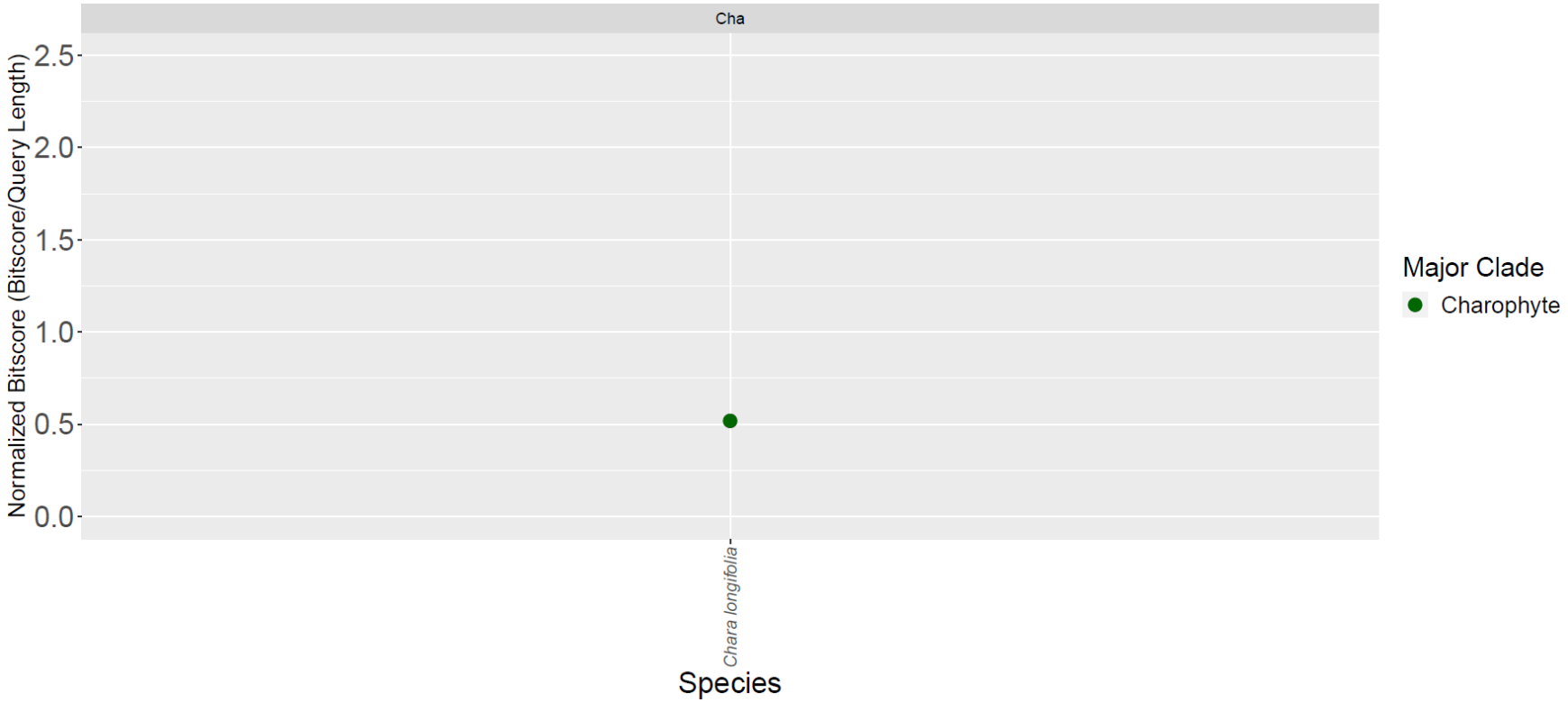
***Spirogyra* Subdomain 2**



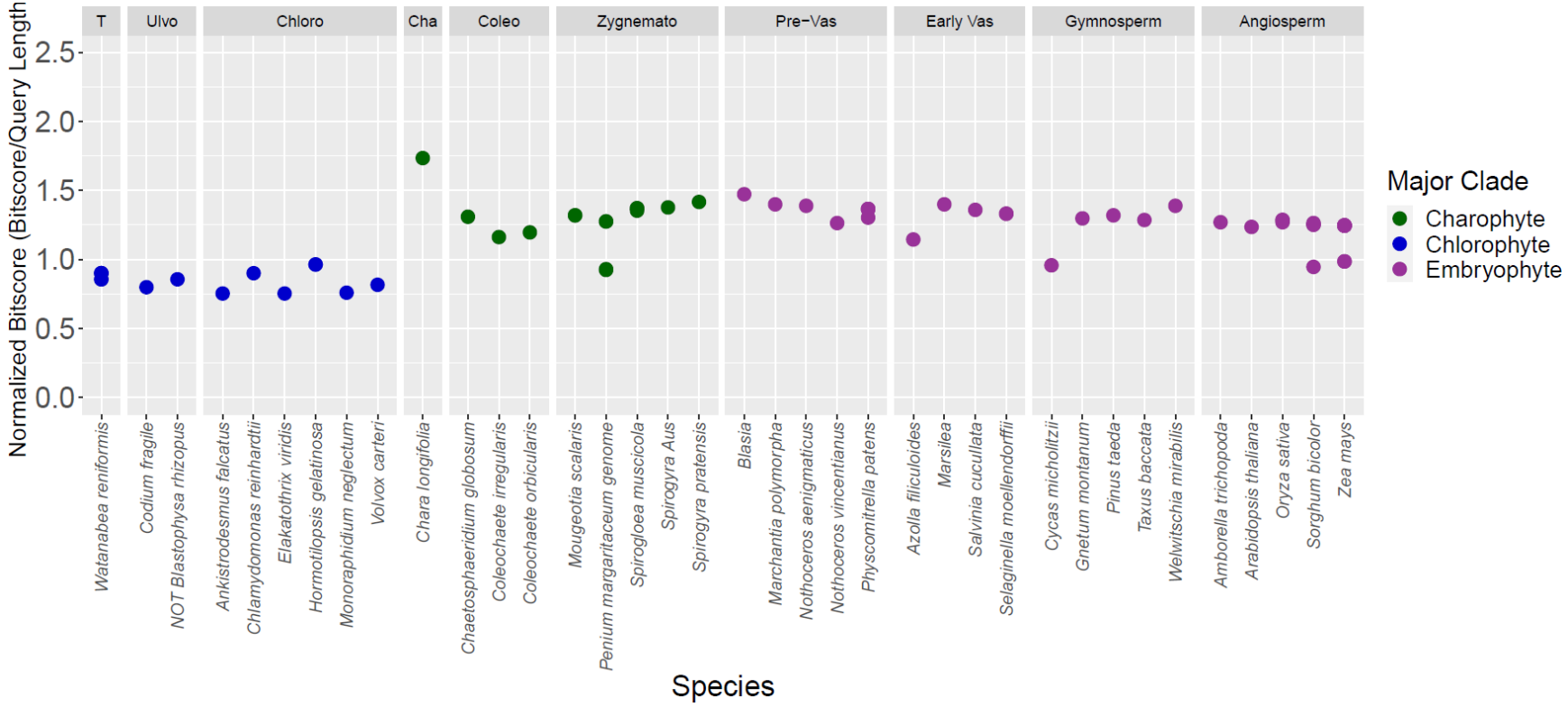
***Spirogyra* Subdomain 3**



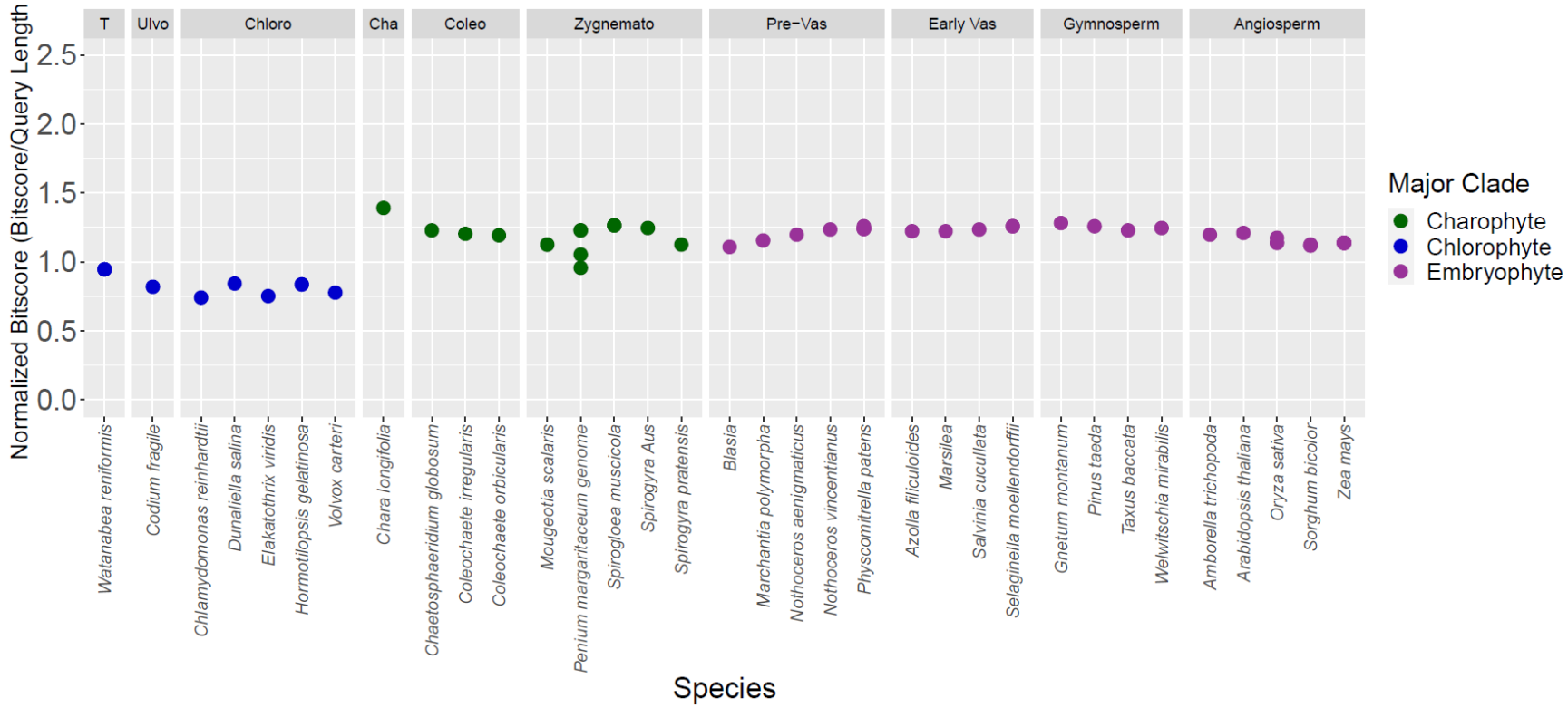
***Nitella* Subdomain 1**



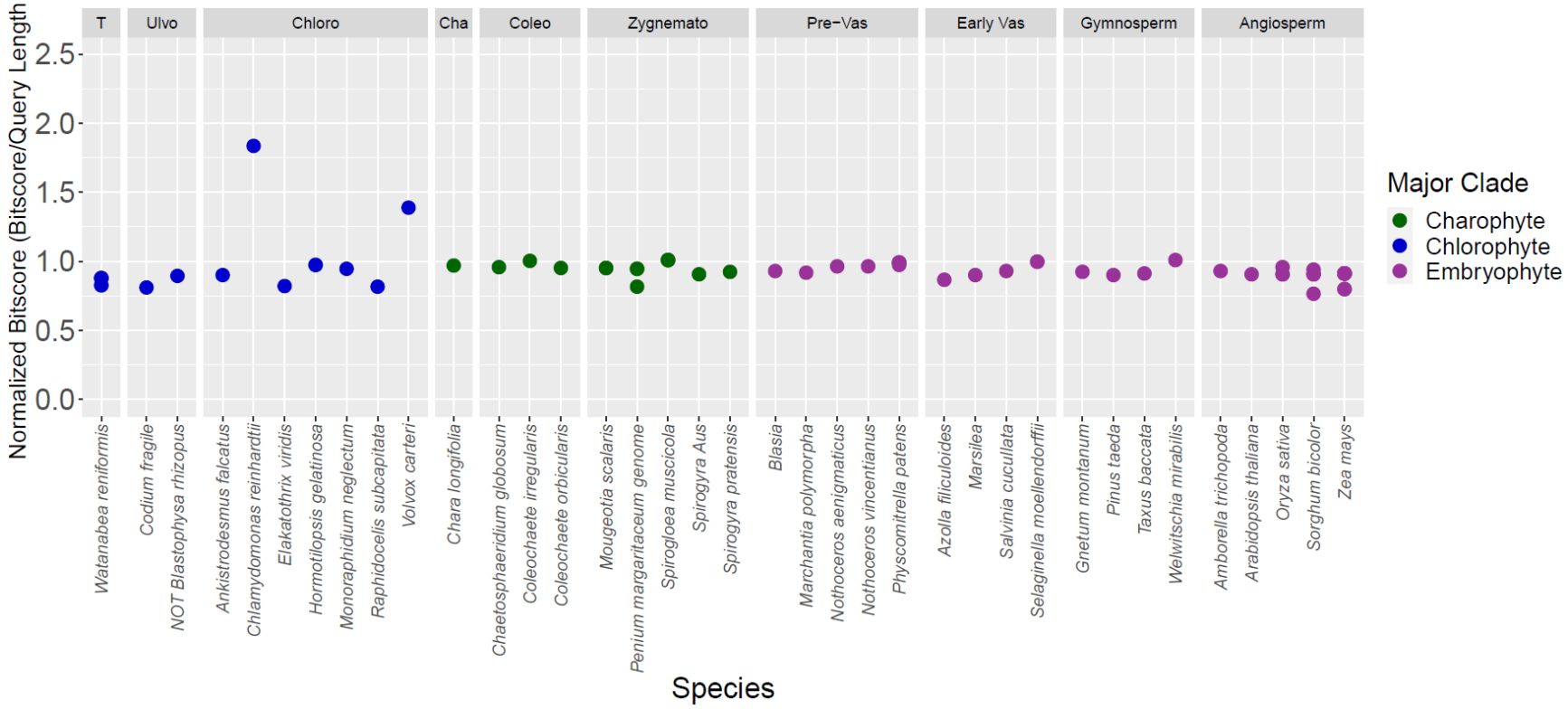
***Nitella* Subdomain 2**



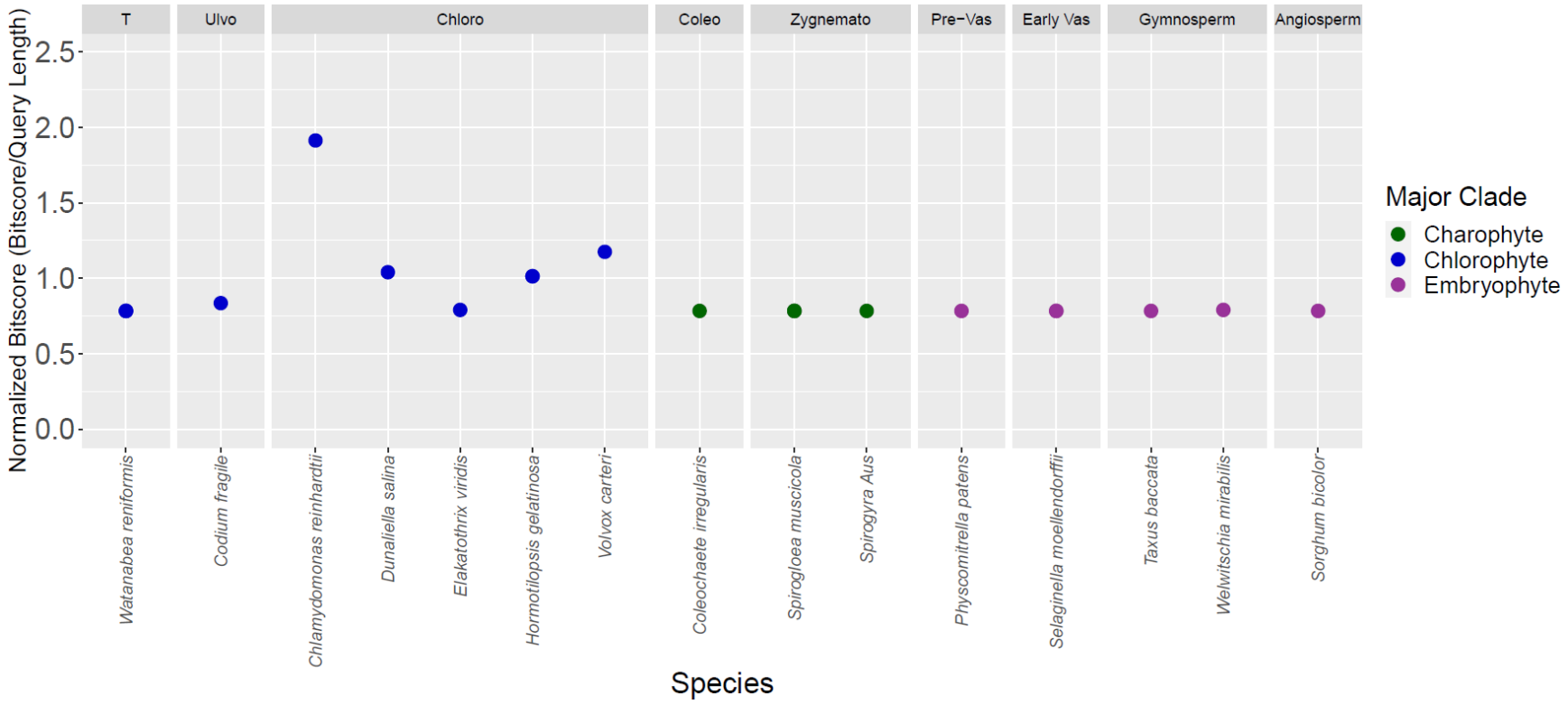
***Nitella* Subdomain 3**



***Chlamydomonas* Subdomain 2**



***Chlamydomonas* Subdomain 3**



Appendix G: Conserved Domains Present in only Two of the Three Major Clades

Embryophytes and Charophytes

cd00066, cd00140, cd00156, cd00195, cd00218, cd00293, cd00294, cd00300, cd00320, cd00382, cd00388, cd00511, cd00525, cd00605, cd00608, cd00631, cd00693, cd00864, cd00871, cd00876, cd00955, cd00958, cd00997, cd01059, cd01131, cd01339, cd01390, cd01555, cd01645, cd01741, cd01745, cd01837, cd01870, cd01874, cd01875, cd01910, cd01953, cd01954, cd01955, cd01956, cd01957, cd02121, cd02142, cd02175, cd02176, cd02220, cd02241, cd02242, cd02243, cd02244, cd02245, cd02262, cd02268, cd02274, cd02284, cd02435, cd02462, cd02475, cd02485, cd02489, cd02604, cd02662, cd02701, cd02705, cd02719, cd02725, cd02870, cd02900, cd02910, cd02930, cd02949, cd02950, cd02985, cd02986, cd03134, cd03169, cd03201, cd03231, cd03238, cd03270, cd03315, cd03319, cd03366, cd03457, cd03527, cd03747, cd03874, cd03905, cd03911, cd03976, cd04019, cd04022, cd04038, cd04105, cd04128, cd04129, cd04130, cd04134, cd04135, cd04145, cd04259, cd04338, cd04360, cd04476, cd04732, cd04878, cd04899, cd04926, cd05035, cd05036, cd05043, cd05045, cd05048, cd05049, cd05050, cd05058, cd05063, cd05065, cd05066, cd05069, cd05072, cd05073, cd05074, cd05079, cd05080, cd05081, cd05085, cd05115, cd05116, cd05199, cd05235, cd05238, cd05290, cd05291, cd05292, cd05293, cd05294, cd05306, cd05318, cd05498, cd05646, cd05804, cd05924, cd05937, cd06040, cd06141, cd06174, cd06269, cd06366, cd06428, cd06591, cd06601, cd06787, cd06825, cd06826, cd06827, cd06899, cd07025, cd07063, cd07204, cd07214, cd07215, cd07431, cd07438, cd07475, cd07504, cd07535, cd07580, cd07607, cd07815, cd07976, cd07990, cd08005, cd08008, cd08022, cd08045, cd08052, cd08060, cd08074, cd08118, cd08165, cd08227,

cd08256, cd08270, cd08377, cd08379, cd08588, cd08603, cd08619, cd08650, cd08662, cd08764, cd08766, cd08875, cd08888, cd08889, cd08890, cd08979, cd09016, cd09098, cd09099, cd09104, cd09137, cd09138, cd09215, cd09218, cd09229, cd09260, cd09261, cd09322, cd09440, cd09441, cd09635, cd09842, cd09843, cd09848, cd10231, cd10312, cd10313, cd10320, cd10538, cd10539, cd10541, cd10543, cd10691, cd10720, cd10851, cd10958, cd11010, cd11028, cd11040, cd11071, cd11074, cd11076, cd11080, cd11084, cd11087, cd11131, cd11187, cd11340, cd11352, cd11593, cd11694, cd11695, cd11698, cd11700, cd11701, cd12125, cd12126, cd12127, cd12129, cd12189, cd12223, cd12233, cd12234, cd12236, cd12237, cd12244, cd12251, cd12298, cd12310, cd12311, cd12334, cd12346, cd12370, cd12389, cd12391, cd12407, cd12454, cd12466, cd12512, cd12524, cd12529, cd12552, cd12576, cd12577, cd12638, cd12640, cd12692, cd12698, cd12771, cd12865, cd12866, cd12868, cd13033, cd13232, cd13396, cd13585, cd13590, cd13686, cd13871, cd13976, cd13987, cd13989, cd13993, cd13998, cd14000, cd14004, cd14024, cd14026, cd14045, cd14053, cd14054, cd14056, cd14057, cd14068, cd14101, cd14107, cd14108, cd14111, cd14112, cd14113, cd14115, cd14146, cd14150, cd14153, cd14156, cd14157, cd14160, cd14174, cd14205, cd14221, cd14222, cd14476, cd14481, cd14507, cd14532, cd14533, cd14535, cd14554, cd14583, cd14584, cd14585, cd14586, cd14587, cd14590, cd14591, cd14592, cd14600, cd14610, cd14614, cd14748, cd14750, cd14767, cd14768, cd14770, cd14819, cd14820, cd14965, cd15028, cd15484, cd15488, cd15662, cd15670, cd15904, cd16013, cd16275, cd16298, cd16343, cd16419, cd17338, cd17406, cd17483, cd17491, cd17546, cd17548, cd17633, cd17636, cd17666, cd17731, cd17781, cd17782, cd17804, cd17893, cd18047, cd18545, cd18546, cd18547, cd18550, cd18563, cd18564, cd18622, cd18802, cd19175, cd19176, cd19180, cd19217, cd19288, cd19305, cd19306,

cd19308, cd19374, cd19494, cd19534, cd19535, cd19544, cd19545, cd19547, cd19597, cd19751, cd19755, cd19907, cd19908, cd19990, cd20068, cd20587, cd20643, cd20644, cd20645, cd20646, cd20647, cd20648, cd20652, cd20653, cd20655, cd20656, cd20657, cd20658, cd20662, cd20663, cd20664, cd20665, cd20666, cd20669, cd20673, cd20674, cd20675, cd20676, cd20677, cd20719, cd20766

Charophytes and Chlorophytes

cd00019, cd00053, cd00054, cd00102, cd00104, cd00113, cd00152, cd00312, cd00322, cd00330, cd00371, cd00391, cd00431, cd00575, cd00576, cd00603, cd00617, cd00716, cd00794, cd00795, cd00842, cd00845, cd01118, cd01327, cd01408, cd01569, cd01677, cd01678, cd01715, cd01752, cd01753, cd01756, cd01757, cd01889, cd01914, cd01951, cd02180, cd02195, cd02318, cd02468, cd02619, cd02621, cd02799, cd02899, cd03000, cd03402, cd03822, cd03893, cd03919, cd03920, cd03921, cd03922, cd03923, cd03924, cd03927, cd03928, cd03929, cd03959, cd03994, cd04109, cd04124, cd04157, cd04199, cd04200, cd04206, cd04222, cd04224, cd04225, cd04226, cd04227, cd04228, cd04229, cd04235, cd04416, cd04515, cd05216, cd05251, cd05272, cd05308, cd05504, cd05509, cd05668, cd05674, cd05675, cd05676, cd05677, cd05680, cd05681, cd06018, cd06030, cd06235, cd06267, cd06275, cd06284, cd06285, cd06290, cd06819, cd06820, cd06821, cd06906, cd06907, cd06908, cd06913, cd07081, cd07121, cd07122, cd07205, cd07225, cd07227, cd07376, cd07381, cd07406, cd07409, cd07565, cd07720, cd07751, cd07901, cd07931, cd07932, cd07993, cd08023, cd08589, cd08942, cd08963, cd09015, cd09164, cd09166, cd09631, cd09763, cd09799, cd09820, cd10325, cd10326, cd10328, cd10329, cd10428, cd10752, cd10771, cd10775, cd11012, cd11018, cd11021, cd11022, cd11023, cd11037, cd11157, cd11166, cd11178, cd11232, cd11300, cd11313, cd11343, cd11355, cd11356, cd11474, cd11477, cd11478, cd11486, cd11487, cd11488, cd11489, cd11490, cd11491, cd11493, cd11494, cd11495, cd11573, cd11574, cd12082, cd12164, cd12180, cd12742, cd12822, cd12828, cd12873, cd12970, cd13022, cd13027, cd13133, cd13617, cd13618, cd13685, cd13687, cd13770, cd14129, cd14130, cd14450, cd14451, cd14452, cd14607, cd14619, cd14790, cd14856, cd15238, cd15241,

cd15902, cd16147, cd16967, cd16968, cd17325, cd17331, cd17334, cd17335, cd17343, cd17345, cd17363, cd17366, cd17367, cd17385, cd17388, cd17390, cd17398, cd17399, cd17438, cd17439, cd17440, cd17488, cd17508, cd17509, cd17769, cd17815, cd17817, cd17842, cd17856, cd17990, cd18025, cd18040, cd18589, cd18600, cd18863, cd18887, cd19226

Embryophytes and Chlorophytes

cd00048, cd00137, cd00326, cd00454, cd00528, cd00649, cd01006, cd01071, cd01094, cd01303, cd01347, cd01516, cd01602, cd01763, cd01801, cd01813, cd01815, cd01846, cd01850, cd01948, cd02048, cd02053, cd02057, cd02059, cd02066, cd02434, cd02497, cd02538, cd03139, cd03378, cd03491, cd03903, cd04048, cd04189, cd04299, cd04708, cd05160, cd05274, cd05280, cd05311, cd05343, cd05346, cd05365, cd05499, cd05687, cd05688, cd06139, cd06173, cd07019, cd07473, cd07477, cd07483, cd07484, cd07489, cd07916, cd08027, cd08092, cd08151, cd08170, cd08171, cd08172, cd08177, cd08186, cd08192, cd08200, cd08242, cd08246, cd08419, cd08420, cd08489, cd08550, cd08623, cd08624, cd08625, cd08626, cd08627, cd08632, cd08934, cd08945, cd08952, cd08953, cd08956, cd09213, cd09597, cd09805, cd10727, cd10734, cd10845, cd11123, cd11350, cd11719, cd12219, cd12376, cd12385, cd12574, cd12617, cd12618, cd13630, cd13633, cd14019, cd14502, cd14656, cd14864, cd14866, cd14949, cd16025, cd16027, cd16034, cd16035, cd16099, cd16104, cd16145, cd16146, cd17061, cd17118, cd17323, cd17327, cd17332, cd17349, cd17355, cd17365, cd17370, cd17538, cd17916, cd18129, cd18790, cd19075, cd19092, cd19138, cd19146, cd19147, cd19218, cd19228, cd19562, cd19567, cd19568, cd19571, cd19752, cd19920, cd19963, cd19984, cd20560, cd20601

References

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A. A., Hoppenrath, M., James, T. Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D. J. G., Lara, E., Le Gall, L., Lynn, D. H., Mann, D. G., Massana, R., Mitchell, E. A. D., Morrow, C., Park, J. S., Pawlowski, J. W., Powell, M. J., Richter, D. J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F. W., Torruella, G., Youssef, N., Zlatogursky, V., & Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119. https://doi.org/10.1111/jeu.12691

Akram, U., Song, Y., Liang, C., Abid, M. A., Askari, M., Myat, A. A., Abbas, M., Malik, W., Ali, Z., Guo, S., Zhang, R., & Meng, Z. (2020). Genome-wide characterization and expression analysis of nhx gene family under salinity stress in gossypium barbadense and its comparison with gossypium hirsutum. *Genes*, *11*(7), 1–22. https://doi.org/10.3390/genes11070803

Bhattacharya, D., & Medlin, L. (1998). Algal phylogeny and the origin of land plants. *Plant Physiology*, *116*(1), 9–15. https://doi.org/10.1104/pp.116.1.9

Blank, C. E. (2013). Origin and early evolution of photosynthetic eukaryotes in freshwater environments: Reinterpreting proterozoic paleobiology and biogeochemical processes in light of trait evolution. *Journal of Phycology*, *49*(6), 1040–1055. https://doi.org/10.1111/jpy.12111

Bowles, A. M. C., Bechtold, U., & Paps, J. (2020). The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. *Current Biology*, *30*(3), 530-536.e2. https://doi.org/10.1016/j.cub.2019.11.090

Brinkmann, H., Van Der Giezen, M., Zhou, Y., De Raucourt, G. P., & Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology*, *54*(5), 743–757. https://doi.org/10.1080/10635150500234609

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9), 1–13. https://doi.org/10.1093/gigascience/giz100

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chanroj, S., Wang, G., Venema, K., Zhang, M. W., Delwiche, C. F., & Sze, H. (2012). Conserved and diversified gene families of monovalent cation/H+ antiporters from algae to flowering plants. *Frontiers in Plant Science*, *3*(FEB), 1–18. https://doi.org/10.3389/fpls.2012.00025

Chen, H. T., Chen, X., Wu, B. Y., Yuan, X. X., Zhang, H. M., Cui, X. Y., & Liu, X. Q. (2015). Whole-genome identification and expression analysis of K+ efflux antiporter (KEA) and Na+/H+ antiporter (NHX) families under abiotic stress in soybean. *Journal of Integrative Agriculture*, *14*(6), 1171–1183. https://doi.org/10.1016/S2095-3119(14)60918-7

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., Wittek, S., Reder, T., Günther, G., Gontcharov, A., Wang, S., Li, L., Liu, X., Wang, J., Yang, H., Xu, Xan., Delaux, Pierre-Marc, Melkonian, B., Wong, G. K., & Melkonian, M. (2019). Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell*, *179*(5), 1057-1067.e14. https://doi.org/10.1016/j.cell.2019.10.019

Cheon, S., Zhang, J., & Park, C. (2020). Is Phylotranscriptomics as Reliable as Phylogenomics? *Molecular Biology and Evolution*, *37*(12), 3672–3683. https://doi.org/10.1093/molbev/msaa181

Cox, C. J., Li, B., Foster, P. G., Embley, T. M., & Civáň, P. (2014). Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology*, *63*(2), 272–279. https://doi.org/10.1093/sysbio/syt109

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., & Von Haeseler, A. (2020). GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Systematic Biology*, *69*(2), 249–264. https://doi.org/10.1093/sysbio/syz051

De Vries, J., Curtis, B. A., Gould, S. B., & Archibald, J. M. (2018). Embryophyte stress signaling evolved in the algal progenitors of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), E3471–E3480. https://doi.org/10.1073/pnas.1719230115

Del Cortona, A., Jackson, C. J., Bucchini, F., Van Bel, M., D'hondt, S., Skaloud, P., Delwiche, C. F., Knoll, A. H., Raven, J. A., Verbruggen, H., Vandepoele, K., De Clerck, O., & Leliaert, F. (2020). Neoproterozoic origin and multiple transitions to macroscopic growth in green seaweeds. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(5), 2551–2559. https://doi.org/10.1073/pnas.1910060117

Delaux, P. M., Radhakrishnan, G. V., Jayaraman, D., Cheema, J., Malbreil, M., Volkening, J. D., Sekimoto, H., Nishiyama, T., Melkonian, M., Pokorny, L., Rothfels, C. J., Sederoff, H. W., Stevenson, D. W., Surek, B., Zhang, Y., Sussman, M. R., Dunand, C., Morris, R. J., Roux, C., Wong, G. K., Oldroyd, G. E. D., & Ane, J. M. (2015). Algal ancestor of land plants was preadapted for symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(43), 13390–13395. https://doi.org/10.1073/pnas.1515426112

Delwiche, C. F., & Cooper, E. D. (2015). The evolutionary origin of a terrestrial flora. *Current Biology*, *25*(19), R899–R910. https://doi.org/10.1016/j.cub.2015.08.029

Delwiche, C. F., Goodman, C. A., & Chang, C. (2017). Land Plant Model Systems Branch Out. *Cell*, *171*(2), 265–266. https://doi.org/10.1016/j.cell.2017.09.036

Delwiche, C. F., & Timme, R. E. (2011). Plants. *Current Biology*, *21*(11), 417–422. https://doi.org/10.1016/j.cub.2011.04.021

Eddy, S. R. (2009). A New Generation of Homology Search Tools Based on Probabilistic Inference. *Genome Informatics*, *23*(1), 205–211.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 1–14. https://doi.org/10.1186/s13059-015-0721-2

Emms, D. M., & Kelly, S. (2018). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *BioRxiv*, 1–14. https://doi.org/10.1101/466201

Finet, C., Timme, R. E., Delwiche, C. F., & Marlétaz, F. (2010). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, *20*(24), 2217–2222. https://doi.org/10.1016/j.cub.2010.11.035

Finet, C., Timme, R. E., Delwiche, C. F., & Marlétaz, F. (2012). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, *22*(15), 1456–1457. https://doi.org/10.1016/j.cub.2012.07.021

Fu, X., Lu, Z., Wei, H., Zhang, J., Yang, X., Wu, A., Ma, L., Kang, M., Lu, J., Wang, H., & Yu, S. (2020). Genome-Wide Identification and Expression Analysis of the NHX (Sodium/Hydrogen Antiporter) Gene Family in Cotton. *Frontiers in Genetics*, *11*(August), 1–16. https://doi.org/10.3389/fgene.2020.00964

Fučíková, K., Leliaert, F., Cooper, E. D., Škaloud, P., D'Hondt, S., De Clerck, O., Gurgel, C. F. D., Lewis, L. A., Lewis, P. O., Lopez-Bautista, J. M., Delwiche, C. F., & Verbruggen, H. (2014). New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Frontiers in Ecology and Evolution*, *2*(OCT), 1–12. https://doi.org/10.3389/fevo.2014.00063

Graham, L. E. (1996). Green algae to land plants: An evolutionary transition. *Journal of Plant Research*, *109*(3), 241–251. https://doi.org/10.1007/bf02344471

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Harholt, J., Moestrup, Ø., & Ulvskov, P. (2016). Why Plants Were Terrestrial from the Beginning. *Trends in Plant Science*, *21*(2), 96–101. https://doi.org/10.1016/j.tplants.2015.11.010

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L., & Hedges, S. B. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science*, *293*(5532), 1129–1133. https://doi.org/10.1126/science.1061457

Hedges, S. B., Blair, J. E., Venturi, M. L., & Shoe, J. L. (2004). A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolutionary Biology*, *4*, 1–9. https://doi.org/10.1186/1471-2148-4-2

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2017). UFBoot2: Improving the ultrafast bootstrap approximation. *BioRxiv*, *35*(2), 518–522. https://doi.org/10.1101/153916

Holzinger, A., & Pichrtová, M. (2016). Abiotic stress tolerance of charophyte green algae: New challenges for omics techniques. *Frontiers in Plant Science*, *7*(MAY2016), 1–17. https://doi.org/10.3389/fpls.2016.00678

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., Moriyama, T., Ikeuchi, M., Watanabe, M., Wada, H., Kobayashi, K., Saito, M., Masuda, T., Sasaki-Sekimoto, Y., Mashiguchi, K., Awai, K., Shimojima, M., Masuda, S., Iwai, M., Nobusawa T., Narise, T., Kondo, S., Saito, H., Sato, R., Murakawa, M., Ihara, Y., Oshima-Yamada, Y., Ohtaka, K., Satoh, M., Sonobe, K., Ishii, M., Ohtani, R., Kanamori-Sato, M., Honoki, R., Miyazaki, D., Mochizuki, H., Umetsu, J., Higashi, K., Shibata, D., Kamiya, Y., Sato, N., Nakamura, Y., Tabata, S., Ida, S., Kurokawa, K,. &Ohta, H. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, *5*(May), 2–6. https://doi.org/10.1038/ncomms4978

Jiao, C., Sørensen, I., Sun, X., Sun, H., Behar, H., Alseekh, S., Philippe, G., Palacio Lopez, K., Sun, L., Reed, R., Jeon, S., Kiyonami, R., Zhang, S., Fernie, A. R., Brumer, H., Domozych, D. S., Fei, Z., & Rose, J. K. C. (2020). The Penium margaritaceum Genome: Hallmarks of

the Origins of Land Plants. *Cell*, *181*(5), 1097-1111.e12. https://doi.org/10.1016/j.cell.2020.04.019

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin, O., Soltis, D. E., Soltis, P. S., Wong, G. K. S., Baker, W. J., & Wickett, N. J. (2019). A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Systematic Biology*, *68*(4), 594–606. https://doi.org/10.1093/sysbio/syy086

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Ju, C., Van De Poel, B., Cooper, E. D., Thierer, J. H., Gibbons, T. R., Delwiche, C. F., & Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. *Nature Plants*, *1*(January). https://doi.org/10.1038/nplants.2014.4

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Karol, K. G., McCourt, R. M., Cimino, M. T., & Delwiche, C. F. (2001). The closest living relatives of land plants. *Science*, *294*(5550), 2351–2353. https://doi.org/10.1126/science.1065156

Kassambara, A. (2016). Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra. Vol. 2. *Sthda*. http://www.sthda.com/english/rpkgs/factoextra

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Laurin-Lemay, S., Brinkmann, H., & Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, *22*(15), R593–R594. https://doi.org/10.1016/j.cub.2012.06.013

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., & De Clerck, O. (2012). Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, *31*(1), 1–46. https://doi.org/10.1080/07352689.2011.615705

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., Depriest, M. S., Bhattacharya, D., Karol, K. G., Fredericq, S., Zechman, F. W., & Lopez-Bautista, J. M. (2016). Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Scientific Reports*, *6*(May), 1–13. https://doi.org/10.1038/srep25367

Leliaert, F., Verbruggen, H., & Zechman, F. W. (2011). Into the deep: New discoveries at the base of the green plant phylogeny. *BioEssays*, *33*(9), 683–692. https://doi.org/10.1002/bies.201100035

Lewis, L. A., & McCourt, R. M. (2004). Green algae and the origin of land plants. *American Journal of Botany*, *91*(10), 1535–1556. https://doi.org/10.3732/ajb.91.10.1535

Mattox, K. R., & Stewart, K. D. (1984). Classification of the green algae: A concept based on comparative cytology. In Irvine, D. E. G., and D. John (eds.), *The systematics of green algae*. Academic Press, London, pp. 29-72.

McCourt, R. M., Delwiche, C. F., & Karol, K. G. (2004). Charophyte algae and land plant origins. *Trends in Ecology and Evolution*, *19*(12), 661–666. https://doi.org/10.1016/j.tree.2004.09.013

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., & Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, *6*(3), 1–15. https://doi.org/10.1002/aps3.1038

Mishler, B. D., & Churchill, S. P. (1985). Transition To a Land Flora: Phylogenetic Relationships of the Green Algae and Bryophytes. *Cladistics*, *1*(4), 305–328. https://doi.org/10.1111/j.1096-0031.1985.tb00431.x

Mishler, B. D., Lewis, L. A., Buchheim, M. A., Renzaglia, K. S., Garbary, D. J., Delwiche, C. F., Zechman, F. W., Kantz, T. S., Chapman, R. L., Mishler, P. B. D., Lewis, L. A., Of, R., Mark, T. H. E., Algae, G., Garbary, D. J., Renzaglia, K. S., Delwiche, B. C. F., Zechman, F. W., Kantz, T. S., & Chapman, R. L. (1994). *Phylogenetic Relationships of the " Green Algae " and " Bryophytes " Source : Annals of the Missouri Botanical Garden , 1994 , Vol . 81 , No . 3 (1994), pp . 451-483 Published by : Missouri Botanical Garden Press Stable URL : https://www.jstor.org/stable/2399900*

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K. K., Haas, F. B., Vanderstraeten, L., Becker, D., Lang, D., Vosolsobě, S., Rombauts, S., Wilhelmsson, P. K. I., Janitza, P., Kern, R., Heyl, A., Rümpler, F., Villalobos, L. I. A. C., Clay, J. M., Skokan, R., Toyoda, A., Suzuki, Y., Kagoshima, H., Schijlen, E., Tajeshwar, N., Catarino, B., Hetherington, A., Saltykova, A., Bonnot, C., Breuninger, H., Symeodi, A., Radhakrishnan, G. V., Nieuwerburgh, F. V., Deforce, D., Chang, C., Karol, K. G., Hedrich, R., Ulvskov, P., Glockner, G., Delwiche, C. F., Petrasek, J., Van de Peer, Y., Friml, J., Beilby, M., Dolan, L., Kohara, Y., Sugano, S., Fujiyama, A., Delaux, P., Quint, M., Theiβen, G., Hagemann, M., Harholt, J., Dunand, C., Zachgo, S., Langdale, J., Maumas, F., Van Der Straeten, D., Gould, S. V., & Rensing, S. A. (2018). The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. *Cell*, *174*(2), 448-464.e24. https://doi.org/10.1016/j.cell.2018.06.033

Proost, S., Pattyn, P., Gerats, T., & Van De Peer, Y. (2011). Journey through the past: 150 million years of plant genome evolution. *Plant Journal*, *66*(1), 58–65. https://doi.org/10.1111/j.1365-313X.2011.04521.x

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Schneider, H., Pisani, D., & Donoghue, P. C. J. (2018). The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, *28*(5), 733-745.e2. https://doi.org/10.1016/j.cub.2018.01.063

Rodríguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B., & Melkonian, M. (2007). Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. *Molecular Biology and Evolution*, *24*(3), 723–731. https://doi.org/10.1093/molbev/msl200

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). *BUSCO: Assessing Genome Assembly and Annotation Completeness BT - Gene Prediction: Methods and Protocols*. https://doi.org/10.1007/978-1-4939-9173-0_14

Strimmer, K., & Von Haeseler, A. (1997). Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 6815–6819. https://doi.org/10.1073/pnas.94.13.6815

Sze, H., Padmanaban, S., Cellier, F., Honys, D., Cheng, N. H., Bock, K. W., Conéjéro, G., Li, X., Twell, D., Ward, J. M., & Hirschi, K. D. (2004). Expression patterns of a novel AtCHX gene family highlight potential roles in osmotic adjustment and K+ homeostasis in pollen development. *Plant Physiology*, *136*(1), 2532–2547. https://doi.org/10.1104/pp.104.046003

Tappan, H. (1980). Charophytes. In: Tappan, H. (Ed.), The Paleobiology of Plant Protists. Freeman and Co, San Francisco, pp. 913–963

Tegeder, M., & Ward, J. M. (2012). Molecular evolution of plant AAP and LHT amino acid transporters. *Frontiers in Plant Science*, *3*(FEB), 1–11. https://doi.org/10.3389/fpls.2012.00021

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE*, *7*(1). https://doi.org/10.1371/journal.pone.0029696

Turmel, M., Otis, C., & Lemieux, C. (2006). The chloroplast genome sequence of Chara vulgaris sheds new light into the closest green algal relatives of land plants. *Molecular Biology and Evolution*, *23*(6), 1324–1338. https://doi.org/10.1093/molbev/msk018

Turmel, M., Otis, C., & Lemieux, C. (2003). The mitochondrial genome of Chara vulgaris: Insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *Plant Cell*, *15*(8), 1888–1903. https://doi.org/10.1105/tpc.013169

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Phillipe, H., dePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K., & Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(45), E4859–E4868. https://doi.org/10.1073/pnas.1323926111

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, *139*(2), 993–1005.

Yarra, R. (2019). The wheat NHX gene family: Potential role in improving salinity stress tolerance of plants. *Plant Gene*, *18*(March), 100178. https://doi.org/10.1016/j.plgene.2019.100178

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(Suppl 6), 15–30. https://doi.org/10.1186/s12859-018-2129-y

Zhong, B., Deusch, O., Goremykin, V. V., Penny, D., Biggs, P. J., Atherton, R. A., Nikiforova, S. V., & Lockhart, P. J. (2011). Systematic error in seed plant phylogenomics. *Genome Biology and Evolution*, *3*(1), 1340–1348. https://doi.org/10.1093/gbe/evr105